applied sciences

MDPI

Article

Automatic Information Extraction from Scientific Publications Based on the Use Case of Additive Manufacturing

Kim Feldhoff [1,*], Hajo Wiemer [1], Philip Träger [2,3], Robert Kühne [2], Martina Zimmermann [2,3] and Steffen Ihlenfeldt [1,4]

[1] Faculty of Mechanical Science and Engineering, Institute of Mechatronic Engineering (IMD), TUD Dresden University of Technology, 01069 Dresden, Germany
[2] Fraunhofer Institute for Material and Beam Technology (IWS), 01277 Dresden, Germany
[3] Institute of Material Sciences (IfWW), TUD Dresden University of Technology, 01069 Dresden, Germany
[4] Fraunhofer Institute for Machine Tools and Forming Technology (IWU), 01187 Dresden, Germany
* Correspondence: kim.feldhoff@tu-dresden.de

Abstract

A systematic literature review is fundamental to building a robust research foundation, informing experimental methodology, and ensuring the quality of future scientific output. However, manual extraction of targeted information from scientific publications is often laborious and prone to error, especially when researchers require rapid access to relevant findings without specialized hardware. This paper introduces an automated workflow for information extraction from scientific publications in the engineering domain. The proposed workflow consists of two primary stages: data preparation and information extraction. During data preparation, PDF files are converted to plain text and segmented into logical sections using a rule-based block detection and classification algorithm for keeping semantics. Information extraction is then performed by applying regular expressions both on keys and values in the same sentence to identify and extract relevant process and material data from the segmented text. The approach was evaluated on a dataset of 18 open-access scientific publications from various journals and conference proceedings in the AM domain. The results of the automated extraction were compared with manual extraction and with a modern large language model (LLM)-based approach. The findings demonstrate that the proposed workflow can accurately and efficiently extract relevant process and material data, achieving competitive performance relative to the LLM-based method. The workflow offers a significant reduction in time and potential errors associated with manual extraction, with automated processing averaging 15 s per document compared to one hour for manual extraction, and achieving a 76% match rate. This efficiency enables researchers to rapidly and effectively extract data. The methodology is readily transferable to other scientific fields where systematic literature reviews and structured data extraction are required.

Keywords: automatic extraction; literature research; scientific publications; information extraction; text mining; PDF format; additive manufacturing

check for updates

Academic Editors: Thomas Lindner and Christos Bouras

Received: 7 July 2025
Revised: 10 August 2025
Accepted: 21 August 2025
Published: 25 August 2025

Citation: Feldhoff, K.; Wiemer, H.; Träger, P.; Kühne, R.; Zimmermann, M.; Ihlenfeldt, S. Automatic Information Extraction from Scientific Publications Based on the Use Case of Additive Manufacturing. *Appl. Sci.* **2025**, *15*, 9331. https://doi.org/10.3390/app15179331

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

1.1. Motivation

The ever-increasing volume of scientific publications presents both a challenge and an opportunity for researchers. Driven by the desire to stay at the forefront of their field and to

make meaningful contributions, researchers are constantly seeking ways to efficiently access and synthesize the wealth of available information. The motivation for this work stems from the need to streamline the research process and to harness technological advances for more effective knowledge discovery.

A systematic literature review is a cornerstone of advancing research, forming a fundamental part of a researcher's knowledge base, experimental methodology, and the quality of future publications. Literature reviews provide an overview of the current state of knowledge on a topic, help gather existing findings, identify gaps in the research, and critically assess methods and results. This enables researchers to avoid duplicating work, develop new research questions, and plan their studies more effectively, thereby laying the foundation for robust, innovative research and enhancing the quality of scientific work. While the objectives of such reviews may vary across disciplines, they universally require effective and efficient information retrieval.

The rapid growth of scientific literature in the field of additive manufacturing (AM) presents significant challenges for researchers seeking to efficiently access and synthesize relevant knowledge. Traditional information retrieval and extraction methods, as surveyed by Chowdhury (2010) [1] and Nasar et al. (2018) [2], provide foundational techniques for indexing and searching scientific texts, yet often lack the domain specificity and automation required for the AM context. Recent advances in deep learning and large language models have enabled more sophisticated extraction and summarization capabilities (Zhu et al., 2020 [3]; Esteva et al., 2021 [4]), but these approaches are frequently constrained by hardware requirements, limited adaptability, and insufficient support for PDF-based scientific documents. There remains a pressing need for algorithms that can automatically and reliably extract domain-specific information—such as experimental setups and parameter ranges—from AM publications, while remaining operable on standard computing resources. This work addresses this gap by proposing a tailored, adaptable solution for automatic information extraction in the AM domain.

1.2. Problem Statement

In contemporary research environments, the volume of scientific publications is rapidly increasing, making quick and efficient access to relevant information more critical than ever. Researchers often rely on manual extraction methods to obtain data from scientific literature, especially from PDF documents. This process is not only labor-intensive and inefficient but is further complicated by the inconsistent structure and formatting of PDFs [5].

Consequently, the process of reviewing and evaluating a substantial number of publications is frequently both time-consuming and unproductive. Furthermore, this approach may result in the identification of irrelevant or inconsequential findings, which do not contribute to the advancement of research or the expansion of the knowledge base. The objective of the AM use case is to extract unique identifiers, such as keywords, in order to identify valuable clusters of information and ranges of physical parameters, thereby providing a basis for comparison. The following section presents the key arguments in favor of implementing an automated extraction tool, primarily for the research-based evaluation of scientific publications.

Additionally, many existing automated solutions require specialized hardware or complex technical setups, which are not always accessible or desirable for researchers. The inability to swiftly extract and process information hinders scientific progress, as researchers spend excessive time on technical tasks rather than on analysis and innovation. This problem affects a broad range of scientific disciplines where literature reviews and data extraction from publications are essential, particularly in fields with rapidly expanding bodies of knowledge. If this challenge remains unaddressed, research workflows will con-

tinue to be slowed down by manual processes and technical barriers, potentially delaying scientific discoveries and reducing research efficiency. There is, therefore, a pressing need for user-friendly, reliable automated tools that enable researchers to extract relevant information from scientific publications quickly and accurately, without the need for specialized hardware or complex configurations.

1.3. Challenges

Automatic information extraction from scientific publications is confronted with a range of challenges arising from both the technical complexity of the data and the practical requirements of end users. To provide a clearer structure, these challenges can be grouped into different clusters. The subsequent sections outline the most important challenges within each cluster, beginning with the cluster "Structural Document" This cluster covers challenges stemming from the diversity of document layouts, formats, and modes of presentation within scientific publications.

Heterogeneity of Document Layouts: Scholarly articles are published in a wide array of layouts, varying by journal, conference, or publisher. This diversity necessitates either highly generalized extraction systems or specialized solutions tailored to specific formats. Developing a universal system demands considerable resources for understanding and adapting to new layouts, potentially reducing efficiency and requiring ongoing maintenance.

Adaptability to Diverse Document Structures: Scientific publications exhibit a wide range of layouts and formatting conventions. The tool must flexibly adapt to these variations, including handling different domain-specific terminologies, to ensure broad usability—especially within additive manufacturing research.

Complexity of PDF Format and Structure: PDF documents, while widely used for their portability, often pose significant challenges for automated information extraction. Unlike structured formats such as XML, PDFs lack a consistent internal structure, making it difficult to preserve semantic relationships during extraction. Hidden layers within PDFs can contain content that is not visible in standard viewers, potentially resulting in incomplete or fragmented extraction of tokens and sentences. Especially with older PDFs created from scanned images, poor image quality may necessitate manual correction—a problem less prevalent in recent publications. To minimize these issues, this work focuses exclusively on modern, well-structured PDFs and excludes image-based files. Nonetheless, the problem of invisible data in hidden layers remains unresolved and is acknowledged as a limitation.

Robust Data Preparation: Accurate extraction relies on effective data preparation. The system must be capable of distinguishing between text, tables, and figures, removing irrelevant layout elements, and normalizing text through methods such as stemming and lemmatization.

In the following, the key challenges related to the cluster "Structural Document" are presented. The key challenges related to the cluster "Linguistic and Semantics" are detailed in the following. This cluster encompasses issues arising from linguistic diversity, terminology, language quality, and the standardization of scientific language.

Variability of Natural Language and Data Representation: A major obstacle is the diverse ways in which information is presented within scientific publications. Although English dominates as the primary language, data can appear in multiple forms—text, tables, figures, or charts. Both the linguistic choices and the format impact the accuracy and reliability of extraction systems. Effective solutions must be able to identify and adapt to these variations either prior to extraction or through robust post-processing mechanisms.

Quality and Clarity of Scientific Publications: The overall quality of scientific writing directly influences extraction performance. Publications authored by non-native speakers may include grammatical or spelling errors, while ambiguous phrasing, misleading content, and complex sentence structures further complicate accurate information extraction. While these issues can be addressed, doing so would significantly increase the initial workload and is beyond the scope of the present work.

Technical Language and Lack of Standardization: The inconsistent use of technical terms, symbols, and units (such as S.I. units) poses a significant challenge. Authors often follow individual conventions, leading to discrepancies that may result in extraction errors or missed parameters. Existing standards like UTF-8 encoding and common scientific symbols can help, but they are not universally applied, making consistent extraction across diverse literature difficult.

The key challenges related to the cluster "Domain-Specific Extraction" are given below. This cluster focuses on the identification and extraction of domain-specific information, parameters, and units tailored to the requirements of specialized research fields.

High Precision in Information Extraction: The utility of the system hinges on the precise extraction of relevant information. This requires the implementation of domain-specific, pattern-based algorithms that can reliably account for the nuances of the target field while maintaining high accuracy.

Necessity for Domain-Specific Expertise: The validation and meaningful interpretation of extracted information demand substantial domain-specific knowledge. In highly specialized fields such as additive manufacturing, materials science, or engineering, the absence of such expertise can hinder the assessment of data accuracy and relevance. Integrating domain knowledge into both the system's extraction logic and its evaluation process is therefore essential to ensure robust and reliable outcomes.

Limitations in Contextual Understanding: Automated extraction systems often struggle to interpret the context in which information appears. Extracted data points, when considered in isolation, can lead to misinterpretation or loss of meaning. While restricting extraction through detailed input parameters can partially address this, additional techniques are required to enhance the system's ability to understand and refine content within its broader textual context.

The key challenges related to the cluster "User-Centric and Accessibility" are presented in the following. This cluster focuses on the need for intuitive, efficient, and widely accessible solutions.

User Accessibility for Non-Experts: Many researchers do not possess advanced skills in text mining or information extraction. Therefore, the system must be highly user-friendly and intuitive, functioning as an out-of-the-box solution that requires minimal configuration while maintaining comprehensive capabilities.

Operation on Standard Hardware: Given that most researchers rely on standard local computers rather than high-performance servers or cloud solutions, the system must be optimized for efficient operation on typical hardware. Achieving this balance between computational efficiency and the demands of document processing is a significant challenge.

Rapid Processing and Timely Results: Researchers depend on timely access to insights from their data. The system must process selected PDF files swiftly, minimizing delays without sacrificing accuracy or the ability to handle large datasets. Algorithmic optimization is key to meeting these expectations.

1.4. Technological Chances

The advent of large language models (LLMs) and AI-based chatbots, such as ChatGPT, has introduced significant opportunities to enhance the literature review process. These technologies provide substantial support in automating the extraction and evaluation of information from scientific publications. For the engineering domain, where the extraction of unique identifiers like keywords and physical parameter ranges is essential, automated tools can identify valuable clusters of information, aiding in comparative analyses and experimental setups.

The automation of information extraction addresses some of the most labor-intensive aspects of literature reviews, such as extracting physical quantities (e.g., tensile strength, elastic modulus) or numerical simulation values for input parameters. This capability is particularly beneficial in the engineering domain, where experimental setups often depend on robust parameter sets derived from existing literature. By automating the extraction process, researchers can accelerate the identification and evaluation of relevant data, reducing the reliance on costly and time-consuming real experiments in favor of virtual simulations.

The automatic extraction of information can be employed to obtain numerical simulation values as input parameters, thereby facilitating the acquisition of target information. The undertaking of real experiments is typically a time-consuming and costly process, which is why the use of virtual experiments is often preferred, and, when feasible, replaced by numerical simulations. As a preliminary step, utilizing pivotal figures from the existing literature as data for simulations could prove invaluable. In particular, in an environment with a limited information base, the use of automatic data extraction could facilitate the acceleration of the detection, procurement, and evaluation process. In the context of AM, the use of parameter sets for a robust manufacturing process could facilitate the replacement of a time-consuming parameter set study through the implementation of statistical experimental designs.

The automatic extraction can make a significant contribution to the acquisition of initial and subsequent impressions of potentially differentiating feature values and data within their contextual environment. The implementation of an automated and documented extraction based on a technological solution will facilitate the reversal and reproducibility of the acquisition and research process, thereby enabling each participant and third parties to comprehend the researcher's methodology.

Additionally, automated extraction contributes to the reproducibility and transparency of research processes. By documenting and systematizing the extraction methodology, the approach ensures that the research process is comprehensible and reversible for third parties, thereby enhancing the reliability and credibility of the findings. In this context, the implementation of an automated extraction tool tailored to AM research is not only a practical necessity but also a significant advancement in the field.

1.5. Main Objective

The primary objective of this work is to develop an algorithm capable of automatically extracting information from scientific publications in PDF format, thereby enabling researchers to rapidly obtain a comprehensive overview of relevant papers—specifically those detailing experimental setups in the engineering domain without the need for specialized hardware, and operable on standard local PCs.

The algorithm must be specifically tailored to the requirements of the engineering domain, ensuring the reliable extraction of domain-specific identifiers such as keywords, clusters of related information, and physical parameter ranges. In addition, the algorithm

must be designed with adaptability as a central consideration, facilitating its extension or modification for application in research domains beyond engineering.

By automating the information extraction process, the algorithm aims to substantially improve the efficiency, accuracy, and productivity of literature reviews, thereby enabling researchers to devote more time to analysis and innovation.

1.6. Research Questions

Based on the challenges identified in Section 1.3, the following research questions are discussed in this contribution:

RQ1: Automated Parameter Extraction: To what extent can parameters, such as physical ranges and material properties, be automatically extracted from scientific publications in the engineering domain?

RQ3: Reliability of Extracted Information: How reliable is the extracted information, and what measures can be implemented to minimize false alarms or irrelevant data? Additionally, how can the system effectively handle and mitigate these inaccuracies?

RQ2: Category-Specific Information Extraction: Is it feasible to automatically extract required information for various predefined categories within scientific publications? What are the essential prerequisites for achieving this objective effectively?

1.7. Outline

The remainder of this work is organized as follows: Section 2 provides a comprehensive overview of existing research and methodologies relevant to the outlined problem of automatic information extraction from scientific publications, with a focus on the engineering domain. The proposed solution is detailed in Section 3, describing the two main components: the data preparation step and the information extraction step. Each step is elaborated with its sub-processes and their role in addressing the outlined challenges. Section 4 presents the application of the proposed approach to the AM use case. It includes a detailed analysis of the results obtained from the automated information extraction process and a comparison with results derived through manual extraction. The advantages and limitations of the proposed approach are critically analyzed in Section 5. Section 6 summarizes the key findings of the work, draws conclusions from the results, and provides an outlook on future research directions and potential improvements to the proposed approach.

2. State of the Art and Related Work

This section provides an overview of the state of the art and related work in text and information extraction from PDF documents, with a particular focus on scientific publications. The exponential growth of research output has made automatic information extraction increasingly essential for efficient, precise, and scalable knowledge discovery. In rapidly evolving fields such as AM, the ability to automatically process and extract relevant information from a vast and heterogeneous literature base is crucial for fostering innovation, accelerating research cycles, and supporting evidence-based decision-making. While both Information Retrieval (IR) and Information Extraction (IE) are foundational for accessing and utilizing scientific knowledge, they serve distinct purposes and employ different methodologies. The following subsections distinguish these concepts and review key advances relevant to automatic information extraction from scientific literature in the AM domain.

2.1. Information Retrieval: Methods and Tools

IR refers to the process of identifying and ranking relevant documents from large collections based on user queries.

Foundational theories in IR are outlined by Chowdhury (2010) [1], who details the principles of document representation, indexing, and relevance ranking. Science mapping, as reviewed by Chen [6], provides systematic methods for visualizing and analyzing the structure and dynamics of scientific knowledge, which is particularly valuable in domains with rapid development and high publication volume, such as AM.

Modern IR systems include both general-purpose search engines (such as Google, Yahoo, and DuckDuckGo) and specialized academic platforms (such as Google Scholar, ScienceDirect, BASE, ResearchGate, and MDPI). These systems enable users to locate publications by querying keywords or phrases, relying on sophisticated indexing, ranking algorithms, and metadata extraction to present a list of potentially relevant documents.

The effectiveness of these search engines is heavily influenced by the quality and breadth of their indexed databases, which can vary significantly among providers. Most search engines operate within a defined corpus, thus limiting the scope of search results to content indexed within their respective ecosystems. For example, Elsevier's ScienceDirect allows users to search for specific identifiers within its proprietary collection, but access is largely restricted to Elsevier's own content and openly available metadata, headers, and abstracts.

As highlighted by Gusenbauer (2019) [7], robust academic search engines such as Google Scholar are essential for effective information retrieval due to their comprehensive coverage and accessibility. However, even these platforms are constrained by the scope of their indexed content and often restrict full-text access because of paywalls or technical limitations. Commercial search engines typically limit public access to metadata—such as titles and abstracts—due to copyright and technical barriers, while freeware search engines generally lack access to protected or commercial databases, further narrowing their utility.

Importantly, while IR systems are invaluable for locating and accessing relevant literature, they do not extract or structure information from within documents. Commercial search engines generally operate on a limited corpus and often restrict public search capabilities to metadata, such as titles and abstracts, due to paywalls and other protective or technical measures that prevent full access to the underlying content. Conversely, freeware search systems typically lack access to commercial and protected databases, which similarly narrows their coverage and utility. As a result, to the best of the authors' knowledge, there are currently no commercial or open-source tools available that enable the automatic extraction, interpretation, and systematic comparison of target information from large numbers of scientific papers in a structured overview.

2.2. Information Extraction: Principles and Approaches

2.2.1. Classical Approaches

Nasar et al. (2018) [2] provide a comprehensive survey of IE techniques, including rule-based, machine learning, and hybrid methods, highlighting their applicability for extracting structured information from unstructured scientific texts. Marcos-Pablos and García-Peñalvo (2020) [8] emphasize the combination of IR with semantic analysis for targeted information extraction, while Wang et al. (2018) [9] demonstrate the successful transfer of specialized IE methods from clinical to other scientific domains, showing the adaptability of these approaches.

1. Rule-Based and Token-Level Methods: Traditional extraction approaches rely on deterministic rules or regular expressions to identify and extract relevant text segments. Python [10] and R [11] offer a range of libraries (e.g., PyPDF2, PdfMiner,

Tabula, PDFQuery, PyMyPDF, Pytesseract) for parsing and extracting text from PDFs. These tools are effective for basic data extraction and pre-processing, such as splitting text into tokens or extracting tables. However, they struggle with complex layouts (e.g., multi-column formats, nested tables, figures with embedded text), and provide little support for semantic interpretation, making them less suitable for nuanced information extraction in scientific literature.

2. Metadata and Content Extraction Tools: Automated tools such as CERMINE [12] and Astera Reportminer [13] are designed to extract metadata (e.g., title, author, affiliations) and structured content from scientific PDFs. CERMINE, for instance, uses machine learning algorithms to segment and classify document components, but is limited to single-document processing. Astera Reportminer, a commercial solution, enables batch extraction and exports data in structured formats (e.g., XLSX, XML), but access is restricted by licensing and proprietary constraints. These tools are valuable for building metadata repositories, but their utility for deep, domain-specific extraction is limited.

2.2.2. Ontology-Based and Semantic Approaches

Semantic approaches, including the use of ontologies and knowledge graphs, facilitate deeper analysis of scientific texts. These methods enable the extraction of entities and relationships based on predefined domain knowledge, thereby improving interpretability and consistency. Gwizdka et al. (2016) [14] introduce Search as Learning (SAL), which conceptualizes search as an iterative, knowledge-building process—an idea highly relevant for extracting and synthesizing AM knowledge. Textpresso [15] leverages ontologies to index and search biological literature, demonstrating the power of structured knowledge representation for IR. Ceci et al. (2012) [16] show how computational ontologies can transform textual data into research networks, facilitating the discovery of relationships and trends in large literature corpora. Dragoni et al. (2012) [17] propose "light ontologies" for improving document and query representation, which can enhance retrieval precision. These approaches are particularly promising for AM, where relationships between materials, processes, and properties are complex and multifaceted.

2.2.3. Deep Learning, NLP, and AI-Based Approaches

Automated information extraction from scientific literature faces substantial challenges: the visually oriented PDF format, the variability of natural language, and the heterogeneity of scientific document structures. To address these limitations, recent advances leverage deep learning and AI-based models. Tools such as GROBID, proposed by Lopez (2009), and pdfminer.six employ machine learning and heuristic methods to reconstruct structured information from PDFs [18,19]. Transformer-based models like BERT and domain-specific adaptations have significantly improved the extraction of entities and semantic relationships from scientific texts, compare Devlin et al. (2019) and Vaswani et al. (2017) [20,21]. Moreover, deep learning techniques now enable more accurate layout and table extraction [22]. Flexible, modular pipelines and systems that learn document structures from annotated corpora further support adaptation to diverse document types. Collectively, these advances enhance the reliability and generalizability of information extraction systems. Ponte and Croft (2017) [23] laid the foundation for probabilistic language models in IR, which have since evolved into deep learning architectures capable of learning complex patterns in text. Zhu et al. (2020) [3] demonstrate the use of deep neural networks to improve retrieval accuracy and handle complex queries, while Zhang et al. [24] apply deep reinforcement learning for dynamic optimization of retrieval processes. Rakhshani et al. highlight the use of automated machine learning (AutoML) for

scalable extraction systems. Tshitoyan et al. (2019) [25] show that unsupervised word embeddings can uncover latent relationships in materials science literature, a technique directly applicable to AM. Esteva et al. (2021) [4] demonstrate the utility of semantic search and abstractive summarization, addressing the need for concise, context-aware synthesis of large document sets.

Tools such as ChatPDF [26] leverage generative AI (e.g., GPT-4.0) for interactive querying and extraction from PDF documents. While these models excel at understanding context and generating human-like responses, current implementations are limited in batch processing, scalability, and domain adaptation. Frameworks by Kovacevic et al. [27] and Dieb et al. [28] combine NLP, AI, rule-based, and machine learning techniques for extracting domain-specific metadata and information. These systems demonstrate high precision and recall, especially when tailored to specific domains such as CRIS or nanocrystal devices. ScienceBeam [29] combines OCR and NLP to extract text and structure from image-based PDFs, enabling the processing of older or scanned documents. Hybrid approaches like this are vital for handling the diversity of formats encountered in scientific publishing.

2.3. Summary of Challenges and Gaps

The scientific literature has produced a diverse set of solutions for the technical barriers in automated information extraction. While no single approach fully resolves all challenges, the combination of advanced PDF parsing, modern NLP, and adaptive extraction pipelines has led to substantial improvements in reliability and generalisability. However, despite substantial progress, automatic information extraction from scientific publications continues to face challenges regarding user accessibility and practical deployment. A major barrier is the lack of user-friendly, out-of-the-box solutions. Most existing tools require programming skills or technical expertise, limiting their accessibility for non-experts. Intuitive interfaces, guided workflows, and comprehensive documentation are often missing, which hampers broader adoption in the scientific community. Hardware requirements present another significant gap. Many advanced systems rely on resource-intensive algorithms, large language models, or cloud infrastructure, making them inaccessible to researchers with only standard local computers. There is a clear need for computationally efficient solutions that operate reliably on typical hardware without complex setup. Furthermore, current methods struggle to handle the diversity and complexity of scientific PDFs, such as multi-column layouts, embedded figures, and inconsistent formatting. Semantic adaptation and domain-specific extraction—particularly for fields like additive manufacturing—remain limited, and many tools are locked behind proprietary platforms, restricting scalability and openness. Finally, timely and responsive processing is often lacking. Many available systems have lengthy runtimes or require manual intervention, which impedes rapid literature review and synthesis. Achieving a balance between processing speed, efficiency, and extraction accuracy—especially when handling large or complex datasets—remains a significant and unresolved challenge for the field. Thus, there is a pronounced need for information extraction tools that are both accessible to non-experts and feasible to operate on standard hardware.

2.4. Conclusion and Outlook

Current methods—including classical information retrieval, ontology-based systems, deep learning, and hybrid approaches—offer a solid methodological basis for automatic information extraction from scientific publications. However, as outlined in the previous section, significant gaps persist, particularly in terms of user accessibility, hardware requirements, adaptability to complex document structures, and the ability to deliver timely results on standard computing resources.

The field is therefore moving towards integrated solutions that combine semantic, AI-driven, and ontology-based techniques while prioritizing user-friendliness and computational efficiency. The future of automatic information extraction, especially in complex and dynamic domains such as additive manufacturing, will rely on the development of adaptable, scalable, and intuitive systems. Addressing these challenges will be crucial for enabling broad adoption, supporting efficient and accurate knowledge discovery, and ultimately empowering researchers to extract meaningful insights from ever-growing scientific literature.

3. Materials and Methods

This section details the materials and methods employed to evaluate the proposed approach for automatic information extraction from scientific publications in the field of AM. The section is structured as follows: Section 3.1 provides an overview of the application context, describes the data basis, outlines the search categories, and defines the reference values used for evaluation. Section 3.2 presents the user input, details the data preparation and extraction steps, describes the underlying algorithm, and discusses the technical implementation. Finally, Section 3.3 explains the procedure for extracting information via the ChatPDF tool using predefined prompts.

3.1. Use Case

This section presents the use case and setup for the information extraction. It includes a detailed delineation of the use case in Section 3.1.1, a comprehensive description of the data foundation in Section 3.1.2, and the definition and explanation of the categories targeted for information extraction in Section 3.1.3.

3.1.1. Description

The use case originates from the collaborative research project AMTwin [30]. Scientists from various research institutions are jointly investigating the interrelationships between manufacturing processes, materials, and material properties using an experimental-numerical methodology. The focus of the investigation is on components produced through the laser powder bed fusion (L-PBF) process, specifically those made from the titanium alloy Ti6Al4V. The aim of the research project was to develop simulation methods for the design of both the manufacturing processes and the manufactured components. Additionally, test methods were devised to enable process monitoring, microstructure characterization, quality testing, and the validation of simulations. Extracting quality-related metrics from the AM-based process chain, which includes the build process via L-PBF, microstructure optimization through heat treatment, static strength evaluation via tensile tests, and fatigue strength assessment through fatigue tests, provides valuable insights into process-structure-property linkages. Moreover, information extracted from the literature offers a systematic overview of existing methods and the range of physical quantities that can act as reference points for defining build parameter values in subsequent or original experiments. The synthesis of information from multiple studies has the potential to aid researchers in identifying commonalities and discrepancies in reported results, which could in turn support the development of more robust experimental protocols and models. A systematic overview that includes key information such as experimental parameters and results can save a significant amount of time and effort compared to manual data extraction, particularly when examining large and diverse textual datasets.

3.1.2. Data Basis

The data basis for this study serves as the foundation for evaluating the applied information extraction methods. This subsection provides an overview of the selection process, characteristics, and structure of the dataset.

Selection Process: The scientific literature on AM has undergone a rigorous selection process to establish the dataset for this study. The articles were manually chosen based on their relevance to the field of AM, with a specific focus on the usage of Ti6Al4V powder. The selection process involved an initial keyword-based search using scientific publication platforms such as ScienceDirect and ResearchGate. This approach resulted in a collection of 18 scientific publications, all published between 2015 and 2018.

Anonymization: To ensure an unbiased analysis, free from the influence of author or publication reputation, and to prevent any correlation between the results of the applied information extraction methods and specific author names or publishers, all publications were preprocessed. In this step, journal and conference names were replaced with single letters, and author names were omitted. While this anonymization makes it difficult to identify individual publications, it is important to note that, in theory, metadata such as the number of pages and figures could still allow for narrowing down the search, provided sufficient computational resources and time are available.

Characteristics: The selected publications are provided as PDF files and originate from a variety of domain-specific journals and conference proceedings. Each publication exhibits unique layouts, such as single-column or two-column per page, and varies in length. Additionally, the proportion of tables and figures differs significantly among the publications. All selected articles are accessible online and are classified as open access. Table 1 summarizes the dataset, detailing aspects such as the number of tables and figures in each publication, as well as the number of columns per page. Figure 1 provides an overview of the number of tokens present in the documents. As shown in the right-hand panel of Figure 1, most documents contain approximately 500 tokens per page. However, four documents exhibit a higher density of textual information, which is attributable to their two-column-per-page layout.

In conclusion, the selected articles represent a diverse and representative sample of scientific publications relevant to a typical literature review of a specific topic, in this case, AM based on Ti6Al4V powder. However, it should be noted that the sample cannot claim to provide comprehensive coverage of the field.

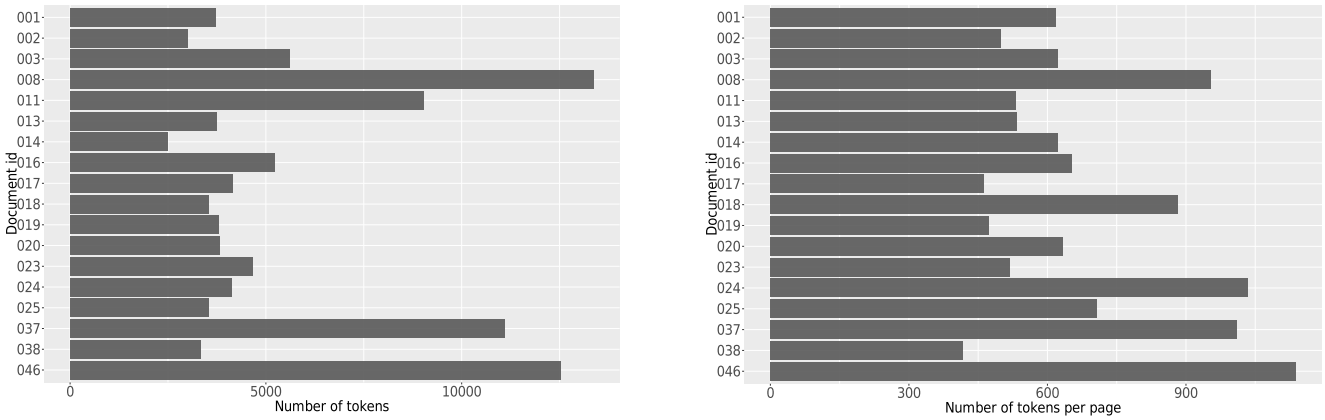

Figure 1. Token counts of selected publications: Total number of tokens across all documents in the dataset. Average number of tokens per page for all documents in the dataset.

Table 1. Summary of the data basis. "Lid": Local document ID; "Gid": Global document ID; "Type": Type of publication (P: conference proceedings, J: journal); "Publisher": Anonymized names of journals and conferences (represented by single letters); "nP": Number of pages in the document; "nC": Number of columns per page in the document; "nW": Total number of tokens in the document; "pW": Average number of tokens per page, defined as pW = nW/nP; "nT": Number of tables in the document; "nF": Number of figures in the document.

Lid	Gid	Type	Publisher	nP	nC	nW	pW	nT	nF
1	001	P	A	6	1	3705	618	2	4
2	002	P	A	6	1	2990	498	2	6
3	003	P	A	9	1	5606	623	1	4
4	008	J	B	14	2	13,367	955	3	18
5	011	J	B	17	2	9039	532	3	21
6	013	P	C	7	1	3733	533	3	5
7	014	P	D	4	1	2491	623	0	6
8	016	J	E	8	1	5219	652	1	5
9	017	J	B	9	1	4162	462	1	8
10	018	P	A	4	2	3529	882	3	6
11	019	J	B	8	1	3793	474	0	7
12	020	P	F	6	1	3806	634	1	4
13	023	P	A	9	1	4671	519	3	8
14	024	P	G	4	2	4135	1034	0	5
15	025	J	B	5	1	3532	706	1	2
16	037	P	A	11	2	11,109	1010	6	9
17	038	P	H	8	1	3331	417	1	7
18	046	J	I	11	2	12,519	1138	4	12

3.1.3. Search Categories

The quality of additively manufactured items is fundamentally influenced by the base material, the configuration of the manufacturing process, and the post-processing through heat treatment [31,32]. Accordingly, the processes involved in powder production, the application of AM using L-PBF, and subsequent heat treatment form the initial primary search categories.

Evaluating the quality of the manufactured items necessitates mechanical testing, particularly tensile testing for analyzing static strength and fatigue testing for assessing cyclic strength. Consequently, the following processes are identified as the primary search category sets:

1. Base material production,
2. manufacturing process,
3. heat treatment,
4. tensile test,
5. fatigue test.

Table 2 outlines the nomenclature for all search categories, providing concise descriptions and enumerating the viable data types for each category.

Table 2. List of category names, along with brief descriptions of each category and the corresponding feasible data types, including string and number.

Id	Search Category Name	Description	Data Type
1	Base_Material_Name	Name of used for base material	String
2	Base_Material_Specification	Specification of base material	String
3	Base_Material_Shape	Shape of base material	String
4	Base_Material_Grain_Size	Grain size in micrometer of base material	Number
5	Base_Material_Supplier_Production	Name of supplier who produced base material	String
6	Manufacturing_Process_Name	Name of manufacturing process	String
7	Manufacturing_Machine	Name of machine used for manufacturing	String
8	Manufacturing_Laser_Power	Laser power in Watt used for manufacturing	Number
9	Manufacturing_Scan_Strategy	Scan strategy used for manufacturing	String
10	Manufacturing_Layer_Thickness	Layer thickness in micrometer used for manufacturing	Number
11	Manufacturing_Energy_Density	Energy density in J/mm^3 used for manufacturing	Number
12	Heat_Treatment_Method	Name of heat treatment method	String
13	Heat_Treatment_Temperature	Temperature in degrees Celsius used during heat treatment	Number
14	Heat_Treatment_Duration	Duration in hours of heat treatment	Number
15	Heat_Treatment_Atmosphere	Atmosphere during heat treatment	String
16	Heat_Treatment_Cooling	Name of cooling type during heat treatment	String
17	Microstructure_Hardness	Measured Vicker's hardness in HV	Number
18	Microstructure_Density_Pores	Measured pore density measure during microstructure test	Number
19	Tensile_Test_Samples_Surface_Condition	Surface condition of samples during tensile test	String
20	Tensile_Test_Youngs_Modulus	Young's modulus in GPa during tensile test	Number
21	Tensile_Test_Yield_Strength	Measured yield strength in MPa in tensile test	Number
22	Tensile_Test_Ultimate_Strength	Measured ultimate strength in MPa in tensile test	Number
23	Tensile_Test_Elongation_At_Failure	Measured elongation at failure in tensile test	Number
24	Fatigue_Samples_Surface_Condition	Surface condition of samples in fatigue test	String
25	Fatigue_Samples_Loading_Direction	Loading direction of test samples in fatigue test	String
26	Fatigue_Test_Bench	Test bench used in fatigue test	String
27	Fatigue_Test_Standard	Standard of fatigue test	String
28	Fatigue_Test_Type_Test	Type of fatigue test	String
29	Fatigue_Test_Loading_Direction	Loading direction in fatigue test	String
30	Fatigue_Test_Frequency	Test frequency in fatigue test	Number
31	Fatigue_Test_R	Stress ratio R in fatigue test	Number
32	Fatigue_Test_NG	Maximum number of cycles in fatigue test	Number
33	Fatigue_Test_Cause_Failure	Cause of failure of sample in fatigue test	String

3.1.4. Reference Values

The following manual information extraction method serves as the reference standard, providing the labelled values that constitute the ground truth. These manually extracted and annotated values are used as a reliable benchmark for evaluating the applied information extraction methods later on. More detailed, the results of the applied methods will be compared against these reference values to assess the accuracy of the methods and to identify errors.

The procedure for manual information extraction from scientific publications is as follows: An expert with specialized knowledge in AM conducts the initial extraction, which is subsequently verified by a second expert in the same field. The involvement of a second expert serves to reduce the risk of missing critical information or recording data inaccurately. Each publication is systematically reviewed by the domain experts using a PDF viewer, with the search function employed to locate specific keywords. The extracted

information is compiled in tabular format within an Excel spreadsheet, where each row represents a document from the dataset and each column corresponds to a search category; each cell thus contains the extracted data for a given document and category as a string.

The results of the manual information extraction method, structured according to the predefined categories, are presented in Figure 2. On average, the manual extraction process required approximately one hour per document. The most time-intensive aspects were the manual keyword search conducted by two domain experts and the subsequent recording of the extracted data in Excel spreadsheets.

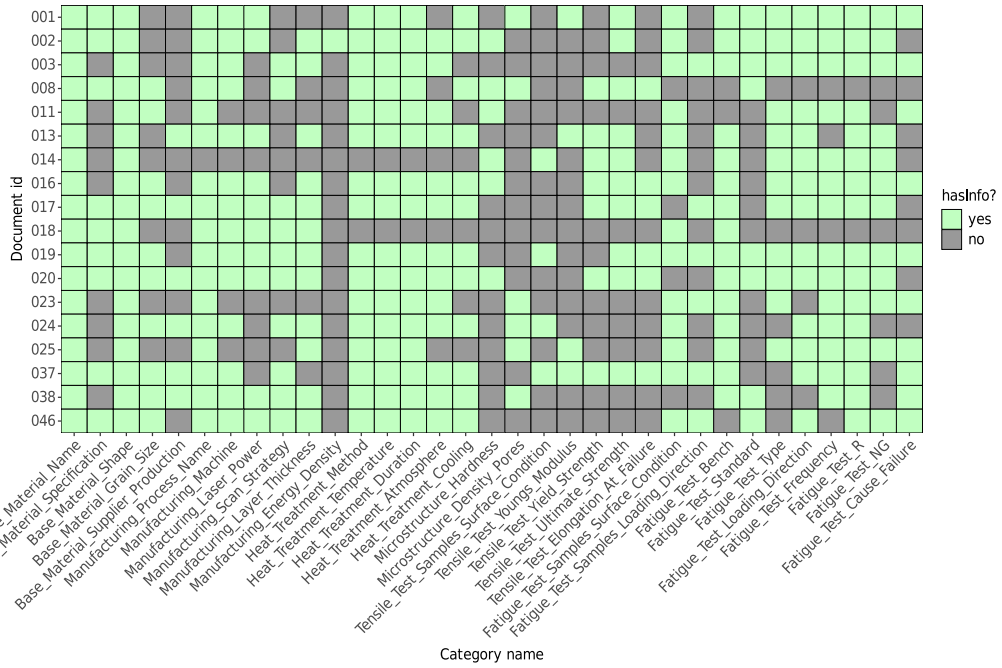

Figure 2. Overview of information for the provided data basis and the selected search categories based on the manual information extraction method "manual" serving as reference standard. In the tabular schema, elements are highlighted in green if the required information is present in the document, and in gray if the information is absent.

Analysis of the manual extraction method given in Figure 2 revealed that categories such as "material name" and "process name" achieved higher detection rates, due to their standardized and prominent reporting in scientific publications. In contrast, categories like "hardness" or "test standard values" showed lower detection rates, because such information is less frequently reported, more context-dependent, or presented in less structured formats. While this observation does not directly affect the results of the applied information extraction methods, it suggests that these approaches could face similar challenges and may exhibit comparable category-specific detection rates. These aspects should be considered when interpreting the robustness and generalisability of the extraction methods.

The manually extracted information also encompasses data from materials containing typographical errors, grammatical inconsistencies, or incorrect physical units. Importantly, the scientific publications in the underlying dataset did not exhibit an unusually high or low frequency of such inconsistencies. As a result, the dataset can be regarded as representative of other domains in terms of the typical occurrence of these types of inconsistencies.

However, the manually created table reveals significant gaps, as information for a substantial number of search categories is not located in the document according to the domain experts. Out of a theoretical maximum of 594 information items (18 publications × 33 categories), only 324 items were identified, resulting in an information detection rate of 54.5%.

It is noteworthy that certain publications, such as those with document IDs 014 and 018, contain only limited information relevant to the selected search categories. Similarly, specific categories—including "hardness," "density," and "pores"—are sparsely populated across the dataset. This pattern can be attributed to several factors. First, some publications may not be closely aligned with the research topic under investigation. Second, certain publications focus exclusively on specific segments of the process chain, such as manufacturing, heat treatment, microstructure analysis, tensile testing, or fatigue testing. Third, confidentiality considerations or insufficient detail provided by the authors may have resulted in limited reporting of configurations and outcomes.

3.2. Automatic Information Extraction

This section outlines the proposed approach for automatic information extraction from scientific publications, specifically tailored to the requirements of AM research. The methodology is structured into two primary components: data preparation and information extraction. The data preparation phase focuses on processing and cleaning the raw textual data extracted from PDF files, ensuring it is suitable for subsequent analysis. The information extraction phase employs targeted algorithms to identify and extract relevant data, such as keywords and parameter ranges, from the prepared dataset.

By systematically addressing the challenges associated with the format and structure of PDF files, as well as the variability in scientific terminology and layouts, this approach aims to streamline the extraction process. The methods described herein are designed to enhance efficiency, reduce manual effort, and deliver high-quality, domain-specific insights. The following subsections provide a detailed description of the data preparation and information extraction steps, along with the tools and techniques employed during their implementation.

The workflow for the automatic information extraction system is structured into three main phases: User Input, Data Preparation, and Information Extraction. Each phase plays a distinct role in ensuring the accurate and efficient extraction of relevant information from scientific publications. Below is a high-level description of the workflow.

User Input Phase: The objective is to define the parameters and criteria for the extraction process. Output is a set of user-defined criteria and patterns guiding the subsequent phases.

Data Preparation Phase: The objective is to process and structure the raw textual data extracted from PDF files into an analyzable format. Output is a clean, structured dataset ready for information extraction.

Information Extraction Phase: The objective is to extract and classify relevant information based on the prepared dataset and user-defined criteria. Output is a structured set of extracted information, ready for analysis and application.

The workflow begins with the user defining the extraction parameters and criteria. This input guides the data preparation phase, where raw textual data are processed into a structured format. Finally, the information extraction phase applies the defined patterns to retrieve and classify relevant data, ensuring accuracy and reliability. This structured approach ensures a robust and efficient system for the automatic extraction of information from scientific publications.

3.2.1. User Input

This section outlines the critical role of user-provided parameters and configurations in guiding the automated information extraction process. Users interact with the system by defining specific search criteria and constraints, which directly influence the accuracy and

efficiency of the extraction process. The following aspects detail the user input requirements and their significance:

Definition of Search Categories: Users are required to provide a set of predefined search categories. These categories specify the types of information to be extracted, such as keywords, physical parameters, or other domain-specific data. Clear definition of these categories ensures that the extraction process targets relevant information, enhancing the precision of results.

Provision of Search Patterns: For each search category, users must supply appropriate search patterns for both keys and values: Keys are strings or patterns representing terms related to the search category (e.g., "Young's modulus"), and values are strings or patterns representing the possible data values associated with the keys (e.g., numerical ranges). Search patterns can be defined using regular expressions, allowing for flexible and robust pattern matching. Logical operators within these expressions enable complex queries, enhancing the system's adaptability to diverse data formats.

Section-Based Search Restriction: Users can optionally restrict the search scope to specific sections of the text, as identified during the data preparation phase. This fine-tuning reduces the algorithm's runtime by focusing only on relevant sections. However, this feature depends on the accuracy of section detection. Misidentified or undetected sections may result in missed key-value pairs, potentially omitting relevant data.

Definition of Numerical Bounds: For search categories involving numerical values, users can define upper and lower bounds to restrict the range of acceptable results. This step ensures that only feasible and reasonable values are considered. By narrowing the search range, this configuration reduces false positives and improves the specificity of the results. For instance, reasonable values for the Young's modulus might be set between [50 GPa; 200 GPa], whereas tensile strength might be constrained to [500 MPa; 1500 MPa].

Users can tailor the extraction process to their specific needs, ensuring relevance and accuracy. By providing detailed input parameters, users can significantly reduce the runtime of the extraction algorithm. Constraints such as numerical bounds and section-based restrictions help minimize false positives and irrelevant results.

The following example demonstrates the user input configuration in a clear and straightforward manner:

1. Search Category: "Layer Thickness": This denotes the type of information the user aims to extract.
2. Search Pattern for Key: "layer", "thick(ness)": These patterns specify the terms related to the search category, allowing the system to identify relevant keywords in the text.
3. Search Pattern for Value: "<floating-point number> <prefix> m": This pattern defines the expected format of the values associated with the key, such as numerical measurements with units (e.g., "0.01 mm").
4. Range for Value: "<floating-point number> <prefix> m": This provides the acceptable range for the numerical values, ensuring only relevant and feasible data are extracted.
5. Search Sections: Introduction, Setup, Results: These are the sections of the document where the system will focus its search, optimizing runtime and improving accuracy by narrowing the scope to relevant content.

The search patterns employed for identifying numerical values, especially those linked to prefixes and physical units, frequently incorporate multiple regular expressions. This inherent complexity poses a challenge for manual validation, as errors such as typographical

mistakes within the expressions can easily occur, potentially compromising the accuracy of the extraction process.

To address the challenges associated with validating and testing regular expressions, it is recommended to utilize both online and offline tools designed for this purpose. Online tools provide a user-friendly interface for quickly testing and refining regular expressions, ensuring their correctness, efficiency, and applicability. These tools are particularly useful for immediate validation needs and offer accessibility from any internet-enabled device. For environments requiring enhanced privacy or when internet access is limited, several offline tools are available across different platforms. On Linux systems, command-line utilities such as grep, sed, and awk offer robust capabilities for regex testing directly in the terminal. For Windows users, PowerShell includes support for regular expressions, and text editors like Notepad++ provide built-in regex functionalities. Additionally, cross-platform tools such as RegexBuddy and integrated development environments (IDEs) like Visual Studio Code or IntelliJ IDEA include comprehensive regex testing features.

Table 3 presents two illustrative examples of regular expressions applied to common search categories. One example showcases a relatively straightforward pattern, while the other demonstrates a more intricate configuration. The complexity of these patterns often increases when accommodating synonyms of the primary technical term. Furthermore, in the context of physical quantities, the regular expressions are crafted to account for potential variations, such as the presence of blanks between numerical values and their corresponding units, and the utilization of different scientific notations for representing numerical values. By considering these factors, the regular expressions are tailored to reliably capture the intended data across diverse textual representations, ensuring robust and precise information extraction.

Table 3. Regular expressions for two frequently used search categories.

Search Category	Feasible Values for Keyword	Regular Expression for Keyword	Feasible Values for Value	Regular Expression for Value
Shape of base material	build, built, produced, producing, manufacturing, manufactured, fabricated, fabricating	(buil(d | t)1,1(d | t)0,1- | produc(ed | ing)0,1- | manufactur(ed | ing)0,1- | fabricat(ed | ing)0,1)	powder, solid, liquid	(powder | solid- | liquid)
Grain size	grain diameter, particle diameter, powder size, particle size, grain size, granulometry	(grain(\s?)diameter- | particle(\s?)diameter- | powder(\s?)size- | particle(\s?)size- | size.*particle- | grain(\s?)size- | granulometry)	“<floating-point number> <prefix> m”	(\d)1,(̇)0,1((\d)0,)- \s?((.)1,1\s?(m)1,1)- ([[\s.,]])0,

3.2.2. Data Preparation

The data preparation phase is a critical component of the proposed solution, laying the groundwork for the subsequent information extraction process. The primary objective of this phase is to create a structured and analyzable textual dataset by extracting and preprocessing textual content from scientific publications. The quality of the dataset produced in this phase directly impacts the effectiveness of the information extraction step. Therefore, meticulous attention is given to ensure that the dataset is clean, well-structured, and ready for analysis. The data preparation phase involves several steps, each addressing specific challenges associated with extracting and structuring textual data from PDF files:

Step 1. Extraction of Textual Data:

1.1. Segmentation of PDF Elements: The content of PDF files is first divided into distinct blocks, separating elements such as tables, figures, text, footers, and captions, compare Algorithm 1. This segmentation is achieved using layout-based indicators, notably vertical spacing, font size, and token positions. This step ensures that the document is segmented into logical units, providing a foundation for subsequent classification and analysis.

1.2. Classification of Detected Blocks: Block types are detected based on layout features such as vertical spacing between lines, font sizes, and the order and position of keywords, compare Algorithm 2. These indicators help in isolating meaningful content from surrounding noise.

Step 2. Text Cleaning:

2.1. Stop Word Removal: Commonly occurring but contextually insignificant words, such as articles and prepositions, are removed to reduce noise and enhance the focus on relevant textual data.

2.2. Removal of Layout-Specific Structures: Elements such as headers, footers, and other repetitive layout-specific structures are eliminated to streamline the dataset.

Step 3. Text Structure Refinement:

3.1. Sentence Decomposition: Text sections are broken down into individual sentences, facilitating easier analysis and processing in subsequent steps.

3.2. Word Normalization: Techniques such as stemming (reducing words to their root forms) and lemmatization (transforming words into their base forms) are applied to standardize the textual data.

Algorithm 1 Algorithm for segmenting each document in a collection into logical text blocks by analyzing layout features such as font height, line spacing, and vertical gaps.

Require: collectionDocs: List of documents (each as a list of pages)
Ensure: collectionDocsBlocks: List of block data for each document

```
for each document in collectionDocs do
    Determine most common font height and line spacing in document
    Compute vertical gap threshold based on these values
    for each page in document do
        Sort tokens by vertical position
        Identify indices where vertical gap threshold
        Define blocks between these indices
        for each block do
            Record position and indices
        end for
        Store block info for page
    end for
    Store all page block info for document
end for
Return collectionDocsBlocks
```

Algorithm 2 Algorithm for classifying text blocks in a document collection by analyzing content, font size, and layout features to automatically identify structural elements such as section headings, figure captions, table captions, headers, and footers

Require: collectionDocs: List of documents
Require: collectionDocsBlocks: List of block layout data
Ensure: Updated block data with classifications
Prepare section name patterns for classification
Get document IDs for all documents
for each document in collectionDocs **do**
  for each page in document **do**
    for each block in page **do**
      Analyze block content, font size, and position
      Detect figure captions (e.g., block starts with "Fig" and font size is unusual)
      Detect table captions (e.g., block starts with "Tab" and font size is unusual)
      Detect section headings (numbering, keywords, width)
      Detect headers (font size, top of page)
      Detect footers (font size, bottom of page, vertical gap)
      Assign block type according to rules and priority
      Update block information
    end for
  end for
end for
Return updated collectionDocsBlocks

To facilitate the detection and categorization of content types within scientific publications, the approach assumes that text is distributed across common sections typically found in scholarly articles. While not all sections need to be present in every document, those that are included must adhere to the following order: Title, Abstract, Keywords, Introduction, Nomenclature, State of the art, Related work (potentially appearing after the Results section), Setup, Experiments/Methods, Results, Conclusion, Acknowledgement, References, Appendix.

This structured assumption ensures that the system can effectively parse and categorize content, even in the absence of certain sections, by relying on the predetermined sequence. This approach not only enhances the robustness of text section detection but also ensures consistency in handling various publication formats.

Step 3 of the data preparation phase focuses on the identification of text sections within scientific publications, leveraging pre-established layout indicators to systematically detect and classify content. This methodology is designed to be adaptable, ensuring it is not constrained by specific journal or conference proceedings templates, thus broadening its applicability across diverse publication formats.

Given the inherent challenges posed by the weak structural definition of the PDF format, the proposed approach employs a custom-developed extraction method. This method prioritizes the preservation of text semantics, ensuring that the extracted content maintains its logical and contextual integrity. Unlike traditional approaches that rely heavily on the physical layout of the page, this method interprets and processes text based on its semantic structure and content, greatly enhancing the accuracy and relevance of the extracted data.

Segmentation of PDF Elements

Step 1 of the data preparation phase focuses on the detection and classification of elements in PDF files as distinct blocks. The block-extraction algorithm given in Algorithm 1 segments a collection of documents into meaningful blocks based on the vertical arrange-

ment of tokens. The process is designed to robustly identify paragraphs or logical units in documents with complex layouts, such as multi-page scientific articles.

- **Document Analysis (lines 1–2):** For each document, the algorithm analyses all tokens across its pages to determine the most common font height and line spacing. These values are used to estimate what constitutes a 'normal' line gap in the document.
- **Threshold Computation (lines 3):** The algorithm computes a threshold for vertical gaps between tokens, typically by multiplying the most common line spacing by a constant factor. This threshold distinguishes between regular line breaks and significant vertical gaps that indicate block boundaries.
- **Page Processing (lines 4–8):** On each page, the algorithm sorts all tokens by their vertical (y) position. It then calculates the vertical distance between consecutive tokens. Whenever a gap exceeds the computed threshold, a new block boundary is defined. The algorithm records the start and end index, as well as the position coordinates, for each block.
- **Result Compilation (lines 9–10):** For each page, the block information is stored in a structured form (such as a data frame or table). All page block data are collected into a list for the document. After processing all documents, the function returns a list of block information for every document in the collection.

Classification of Detected Blocks

The goal of the algorithm given in Algorithm 2 is to assign semantic types to text blocks in a collection of documents based on their content, font, and layout features. The process operates at the document, page, and block levels.

- **Preparation (lines 1–2 of Algorithm 2):** The algorithm begins by preparing a set of section name patterns (such as "Abstract", "Introduction", etc.) to assist in identifying section headings. It also retrieves unique document IDs for reference and verification.
- **Document and Page Iteration (lines 3–4):** For each document in the collection, and for each page within that document, the algorithm iterates through all detected blocks. Each block is initially assigned the default type 'Text'.
- **Feature Extraction (lines 5–6):** For each block, the algorithm extracts:
  - The block's text content (first and last token)
  - Font size and vertical position (y-coordinates)
  - Block width (difference between leftmost and rightmost x-coordinates)
  - The number of tokens in the block
- **Block Type Detection (lines 7–13):** Several rules are applied in order of priority:
  - **Header:** If the block uses a different font size and is positioned at the top of the page, it is classified as a header.
  - **Footer:** If the block uses a different font size, is at the bottom of the page, and is separated by a large vertical gap from the previous block, it is classified as a footer.
  - **Figure/Table Caption:** If the block starts with "Fig" or "Tab" and has an unusual font size, it is classified as a figure or table caption, respectively.
  - **Section Heading:** If the block starts with a section keyword or a numbered pattern and is relatively narrow (not extending to the end of the column), it is classified as a section heading.
  - **Text:** If none of the above conditions are met, the block remains classified as regular text.

  The order of these checks is crucial, as a block may match several criteria, but only the first matching type is assigned.

- **Update and Output (line 17):** The determined type is assigned to the block in the output structure. After all blocks are processed, the updated classification is returned for the entire document collection.

Document- and Token-Level Data Structures for Scientific Texts

A central component of the data preparation phase is the creation of a clean, structured dataset. Within this context, a clean, structured dataset refers to a collection of textual data that has been systematically segmented, annotated, and organized according to clear rules and standards. To create such a dataset, a structured and methodical approach is employed:

1. **Document-Level Metadata Extraction:** For each scientific publication, key metadata such as document ID, title, authors, creation and modification dates, and subject areas are extracted. These data points are organized in a dedicated metadata table (see Table 4), enabling easy identification, filtering, and referencing of documents.
2. **Text Tokenization and Annotation:** The full text of each publication is processed and segmented into logical blocks, sentences, and tokens. Each token is annotated with linguistic features (lemma, part-of-speech, named entity) and positional information. This information is stored in a token-level table (see Table 5), with each entry referencing the corresponding document via a shared Doc ID.
3. **Relational Linking:** The two tables are linked by the Doc ID, ensuring that every token or sentence can be traced directly to its parent document and associated metadata. This relational structure supports robust querying and cross-referencing.

This way of storing and structuring has the following advantages:

Clarity and Reproducibility: The separation of metadata and token-level information provides clear data organization and supports reproducible workflows.
Scalability: The relational design allows for efficient storage and processing, even with large document collections.
Analytical Flexibility: Researchers can perform analyses at both document and token level, enabling a wide range of text mining and information extraction tasks.
Traceability: The Doc ID linkage ensures that all data points remain connected to their original context, supporting transparency and data integrity.

Table 4. Data documentation for the metadata table used in the data preparation phase. Each column is described in terms of its name, data type, purpose within the dataset, and a representative example value.

Column	Data Type	Description	Example Value
Doc ID	String	Unique document identifier	002-AM
DOI	String	Digital Object Identifier for the document	10.1038/nature12345
Filename	String	Name of the PDF file	002-AM.pdf
Title	String	Title of the document	Climate Impact on Urban Areas
Author	String	Authors of the document	Jane Smith; John Brown
Pages	Integer	Number of pages in the document	12
Modified	DateTime	Date and time of last modification (YYYY-MM-DD hh:mm:ss)	2021-05-14 10:23:45
Created	DateTime	Date and time of creation (YYYY-MM-DD hh:mm:ss)	2021-05-10 09:00:00
Creator	String	Software or person who created the document	LaTeX (TeX Live 2021)
Producer	String	Software that produced the PDF	PDFTeX-1.40.21
Keywords	String	Keywords related to the document	climate, urban, change
Subjects	String	Subject areas of the document	Environmental Science

Table 5. Specification of the token-level data structure for efficient storage and processing of scientific texts. Each column is defined by its name, data type, purpose, and an illustrative example.

Column	Data Type	Description	Example Value
Doc ID	String	Foreign key linking to the document in the metadata table	002-AM
Block ID	Integer	Identifier for a logical text block within the document	1
Sentence ID	Integer	Identifier for a sentence within a block	1
Token ID	Integer	Identifier for a token (word or punctuation) within a sentence	1
Token	String	Surface form of the token as it appears in the text	Available
Lemma	String	Lemma or base form of the token	available
POS	String	Part-of-speech tag (e.g., NOUN, VERB, ADJ)	ADJ
Entity	String	Named entity tag (if present) or empty if not applicable	
Section ID	String	Identifier for the section of the document (e.g., Introduction, Methods)	Prolog
TypeBlock	String	Type of text block (e.g., paragraph, title, table)	1-Text
Page	Integer	Page number in the PDF where the token is located	1
IdxGlobalStart	Integer	Global character offset for the start of the token in the document	1
IdxGlobalEnd	Integer	Global character offset for the end of the token in the document	10

Limitations and Requirements for PDF Processing

The proposed solution is designed with specific requirements and limitations regarding the processing of PDF files. A key limitation is that the approach does not support the processing of PDF files that have been converted into image-based formats. The methodology relies on the extraction of textual information in plain text format, which necessitates that the PDF files retain their text-based structure.

- **Handling Multiple Layers in PDF Files:** PDF files often contain multiple content layers, with only one layer typically visible in standard PDF viewers. To address this, the proposed approach requires the flattening of PDF files prior to processing. Flattening ensures that hidden layers are merged and their content becomes accessible for extraction. However, this process is effective only if the PDF files have not already been flattened into a single layer during their creation or prior modifications. If the original layered structure is lost, some information might remain inaccessible.
- **Importance of Block Detection and Classification:** The processing workflow includes two critical steps: block detection and block type classification. Of these, block detection is particularly crucial, as it establishes the foundation for subsequent analyses. If the detected blocks do not align with the semantic requirements of the system, the subsequent information extraction process, which relies heavily on the accurate segmentation and extraction of sentences, may fail. Ensuring the precise detection of text blocks is therefore essential to maintaining the operability and accuracy of the entire extraction pipeline.

3.2.3. Information Extraction

Following the data pre-processing phase, as detailed in the preceding section, the next step involves the extraction of information from the prepared textual dataset. This phase focuses on identifying and classifying relevant data, specifically keywords and value ranges of physical quantities, to enable their structured analysis and application in AM research. The primary objectives of this phase are:

Keyword Extraction: The process involves identifying terms and phrases that are specific to the field of AM research. Additionally, it accounts for variations in how keywords are represented, addressing challenges such as typographical errors and inconsistencies in terminology.

Value Range Classification: The process includes extracting numerical values along with their associated units, such as physical parameters, directly from the text. Furthermore, it ensures the accurate classification of these values, even in the presence of discrepancies in unit representation or formatting errors.

A major challenge in this phase is the variability in how keywords and numerical values are presented. Differences in typographical conventions, character usage for units, and formatting inconsistencies must be addressed by the algorithm to ensure reliable extraction and classification.

The extraction process is performed using a combination of search patterns tailored to each search category. The methodology is as follows:

Pattern Matching: Specific patterns are defined to identify keywords and numerical values within the text. These patterns are designed to accommodate variations in spelling, formatting, and unit representation.

Category-Specific Processing: For keywords, the algorithm uses a predefined list of domain-specific terms, supplemented by contextual analysis to identify relevant additions. For value ranges, the algorithm identifies numerical data and validates associated units, ensuring consistency and relevance to the AM domain.

Error Handling and Refinement: Mechanisms are implemented to detect and correct errors in extracted data, such as mismatched units or incomplete numerical entries. The system iteratively refines its output by cross-referencing extracted data with predefined standards and expected formats.

By employing a structured combination of search patterns and category-specific algorithms, the information extraction phase ensures the reliable identification of keywords and numerical values.

Once this task has been completed, the extraction of information from the text dataset can be performed using straightforward methods. The methodology employed here is a combination of suitable search patterns for each search category.

3.2.4. Algorithm

The algorithm given in Algorithm 3 outlines the process used for automatic information extraction within the defined workflow. It iterates through the entire dataset, systematically searching for key-value pairs within the text sections of scientific publications. The algorithm is designed to ensure that both the key and its corresponding value are identified within the same sentence, which is a practical assumption for most structured scientific texts. Additionally, the algorithm incorporates validation for numerical values based on predefined ranges, enhancing the reliability of the extracted data.

The algorithm produces a structured list containing all identified keyword-value pairs for each sentence within the specified search categories. Each entry in the list includes the sentence identifier (Sentence ID), the corresponding keyword and value, and the section of the document where the sentence is located. This output provides a comprehensive mapping of extracted information, facilitating further analysis and ensuring traceability within the document.

Although the present study evaluates the algorithm solely within the context of AM, the method has been intentionally developed for adaptability. The algorithm utilizes externally defined, configurable search categories and regular expression patterns, ensuring it is not intrinsically tied to the specific terminology or structure of AM literature. Consequently, the algorithm can, in principle, be readily adapted for use in other research domains by modifying the category definitions and extraction criteria to suit the requirements of the target field.

Algorithm 3 Algorithm for automatic extraction of key-value pairs from prepared scientific text sections using user-defined patterns for keys and values, with optional numerical value range constraints, and storage of results together with sentence identifiers.

Require: *SearchCategories*: List of search categories.
Require: *KeyPattern*: User-defined patterns for identifying keys in each category.
Require: *ValuePattern*: User-defined patterns for identifying values in each category.
Require: *TextSections*: Prepared text sections containing sentences.
Require: *ValueRanges*: Optional ranges for numerical search categories.
Ensure: *Results*: Collection of extracted key-value pairs with sentence identifiers.

```
for all category C ∈ SearchCategories do
  for all section S ∈ TextSections do
    for all sentence T ∈ S do
      if KeyPattern(C) and ValuePattern(C) are found in T and ValueRanges(C) is satisfied then
        Store T's identifier (SentenceID) along with the extracted key and value.
      end if
    end for
  end for
end for
return Results
```

3.2.5. Implementation

The implementation of the proposed methodology has been carried out using R [33], a robust and widely used open-source software environment for statistical computing and data analysis.

To ensure maximum reproducibility, all components required for executing the tests have been encapsulated within a Docker container image. This approach provides a controlled and consistent execution environment, facilitating the replication of results. The Docker container image is based on the lightweight Alpine Linux operating system, selected for its minimal resource requirements and efficiency. It includes an installation of R, along with all necessary R packages and self-developed R scripts required for performing the tasks outlined in the methodology. By isolating the environment, this setup ensures that the tests can be executed under identical conditions, independent of the host system's configuration.

Additionally, the use of the containerization approach significantly reduces the risk of side effects. Unlike traditional software installations that may include extraneous components, the container image is streamlined, containing only the essential software and dependencies required for the execution of the specified tasks. This minimization not only enhances performance but also simplifies debugging and maintenance processes.

Moreover, the containerized implementation supports scalability and portability. The encapsulated image can be deployed across various systems, ensuring that the methodology can be utilized and tested in diverse computational environments without compatibility issues. This makes the approach particularly suitable for collaborative research projects, where consistency and reproducibility are critical.

All software necessary for conducting the tests, including custom-developed R scripts, was packaged within a Docker container. This container was then executed on a virtual machine equipped with four cores and 16 GB RAM, running Ubuntu 22.04 LTS. As a result, this setup closely resembles that of a typical modern local PC.

3.3. Information Extraction Using ChatPDF

This section describes the application of ChatPDF for information extraction from scientific publications. ChatPDF is used here as a representative example of AI- and LLM-based extraction methods.

For the analysis, each scientific publication—available as a PDF—was uploaded individually to the ChatPDF web interface, with a separate chat session initiated per document. The queries, formulated as ChatPDF prompts, are detailed in Listing 1. ChatPDF utilizes smart dynamic routing to process queries, employing either the GPT-4o or GPT-4o-mini large language models. Once ChatPDF generated a response, the answer was manually copied from the web interface and saved as a text file.

Listing 1. Query in ChatPDF used for extracting information from each document of the data basis and each search category.

```
1 What are the values of the following quantities? Please arrange the found
      values in a tabular schema.
2 The name, the specification, the shape, the grain size, and the name of
    the supplier of the base material.
3 The name of the manufacturing process, the name of the machine, the laser
      powder, the scan strategy, the layer thickness, and the energy
    density used during the manufacturing process.
4 The name of the heat treatment method, the temperature during the heat
    treatment, the duration of the heat treatment, the atmosphere used
    during the heat treatment, the cooling method after the heat
    treatment method.
5 The microstructure hardness, the density of the pores within the
    microstructure.
6 The surface conditions of the test samples used in the tensile tests, the
      Young's modulus E in GPa, the Yield strengths R_e in MPa, the
    ultimate strengths R_m in MPa, and the elongations at failure A in
    percent of the test samples in the tensile tests.
7 The surface conditions, the loading directions of the test samples used
    in the fatigue tests, the test bench and the test parameters used in
    the fatigue tests, the test standards in the fatigue tests, the types
      of fatigue tests, the loading direction of fatigue tests, the test
    frequency, the ratio R used during the fatigue tests, the maximum
    number of cycles N_G achieved during the fatigue tests, the cause of
    failure of samples in fatigue tests.
```

4. Results

This section contains the results from the application of the proposed automated workflow combining data preparation and algorithmic information extraction ("automatic"), and the information extraction using defined prompts with the ChatPDF tool ("chatPdf").

4.1. Method Automatic

The following subsections present the results of the proposed automatic information extraction method "automatic". Section 4.1.1 addresses the outcomes of the data preparation phase, including the own block detection and classification algorithm for preserving text semantics, while Section 4.1.2 shows the results of the information extraction phase, in which regular expressions are used to identify keys and values.

4.1.1. Data Preparation

Figure 3 provides an overview of the various block types that were identified, with examples drawn from a scientific publication related to AM. Furthermore, the approach takes into account the information presented in tables and figures. The extraction of elements containing figures is conducted in an automated manner. In considering the textual information present within figures, the sentences will be kept relatively short.

The accuracy of block detection and subsequent block type classification is largely influenced by the layout of the publications being processed. Figure 3 depicts an example of a publication with a single-column layout per page. This straightforward layout facilitates accurate identification and classification of blocks. However, in cases involving more complex layouts, the likelihood of errors in block detection and classification increases. Overlaps

between blocks are most commonly observed in instances where tables or figures contain annotated text. Despite these challenges, the results achieved are generally satisfactory.

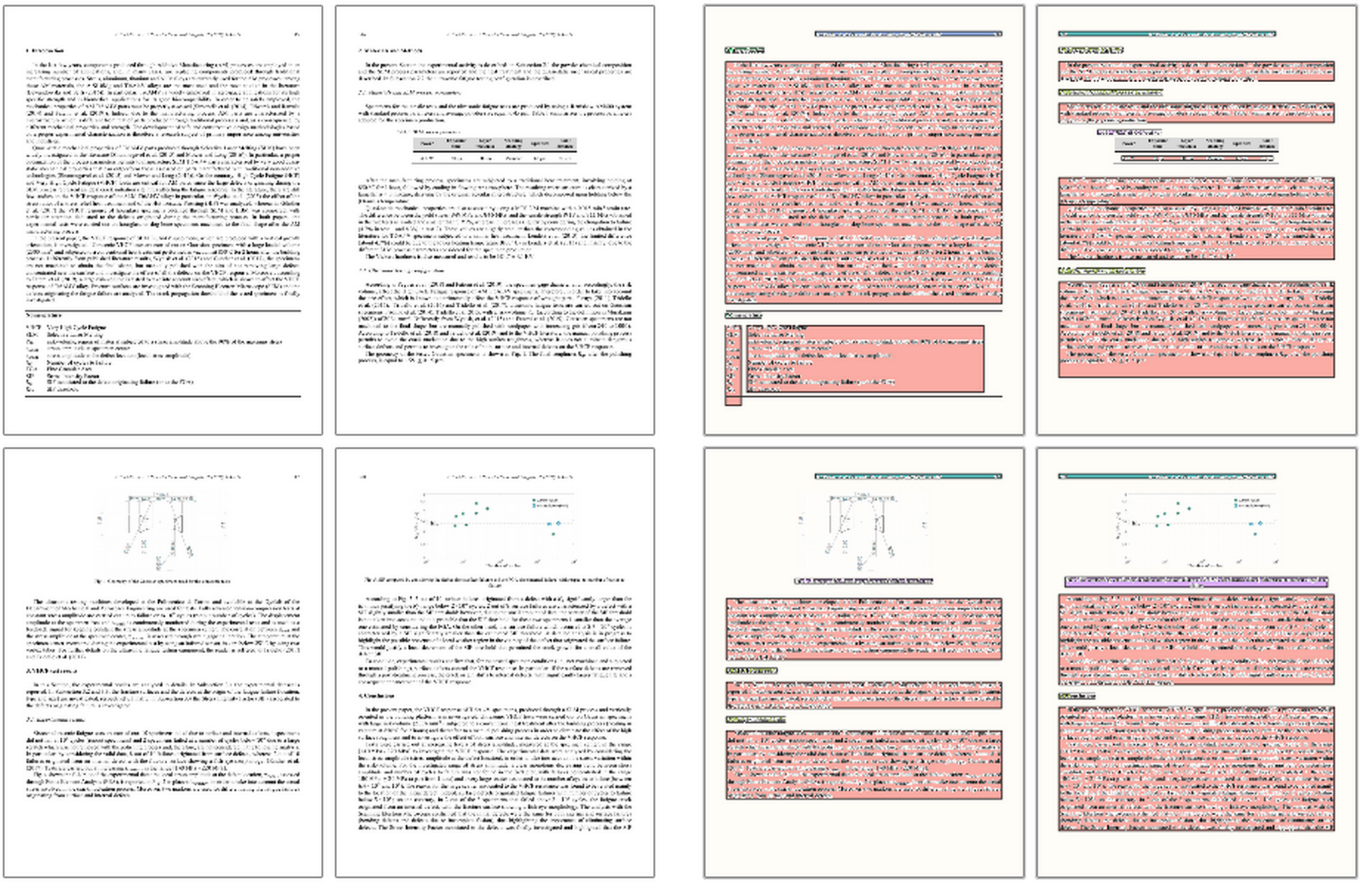

Figure 3. Thumbnails of selected pages from a scientific publication pertaining to AM are presented. To ensure the anonymity of the specific publication, thumbnails with unreadable text are utilized. The left figure shows the original thumbnails, while the right figure displays the thumbnails with detected blocks. The blocks are color-coded according to their classifications: red for text blocks, green for header blocks, and purple for section headers.

4.1.2. Information Extraction

The following presents the outcomes of the proposed automatic information extraction method, structured according to the predefined categories. The method relies entirely on rule-based and deterministic algorithms. As a result, only a single run is required, and there are no hyperparameters to be tuned. All necessary information for running the tests is contained within a configuration file, which specifies the regular expressions for each search category. Each category is already equipped with several regular expressions combined to achieve optimal performance.

To illustrate how values were extracted by the proposed method, Table 6 lists the extracted values for the category Manufacturing_Layer_Thickness across all documents in the dataset. Relevant details, such as the section name and sentence ID, are also provided to indicate where these values can be found within the documents.

The tabular schema given in Figure 4 illustrates the results of the proposed extraction approach for each combination of search category (displayed on the x-axis) and document (listed on the y-axis).

Table 6. Values for the category Manufacturing_Layer_Thickness extracted by the proposed method "automatic". "Local ID": Local document ID; "Global ID": Global document ID; "Values": Detected information, "NA" indicates that values were not found or are not available in the document; "Keys": Corresponding keys found in the same sentence as the values; "Section": Names of section within the document in which extracted information is located; "Sentence IDs": Identifiers of sentences in which extracted information has been found.

Local ID	Global ID	Values	Keys	Sections	Sentence IDs
1	001	NA			
2	002	30 µm	layer, thickness	Setup	65
3	003	50 µm	layer, thickness	Approach	120
4	008	NA			
5	011	NA			
6	013	NA			
7	014	NA			
8	016	NA			
9	017	60 µm	layer, thickness	Setup	89
10	018	NA			
11	019	60 µm	layer, thickness	Setup	55
12	020	60 µm	layer, thickness	Approach	79
13	023	NA			
14	024	100 µm	thickness	Approach	128
15	025	60 µm	layer, thickness	Setup	85
16	037	NA			
17	038	NA			
18	046	50 µm	layer, thickness	Nomenclatur	129

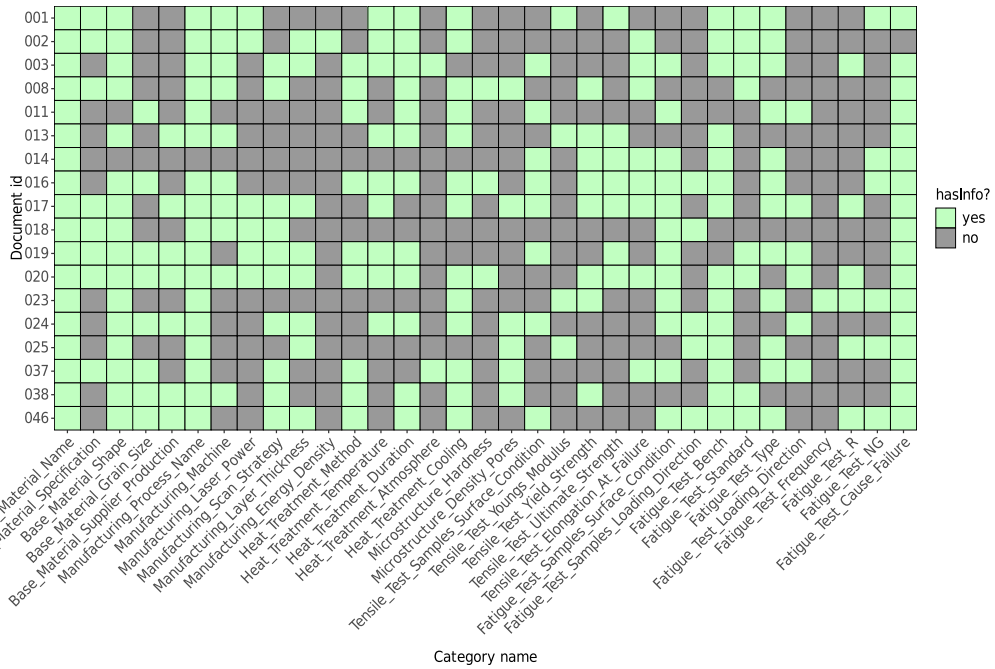

Figure 4. Overview of the detected information for the provided data basis and the chosen search categories using the proposed information extraction algorithm "automatic". In the tabular schema, elements are highlighted in green if information was successfully detected, and in gray if no information could be identified.

4.2. Method chatPdf

In the following, the outcomes of the information extraction method "chatPdf" are presented.

To illustrate how the values were extracted by the proposed method, Table 7 presents the extracted values for the selected search categories for one document in the dataset.

The tabular schema given in Figure 5 illustrates the results of the extraction method for each combination of search category (displayed on the x-axis) and document (listed on the y-axis).

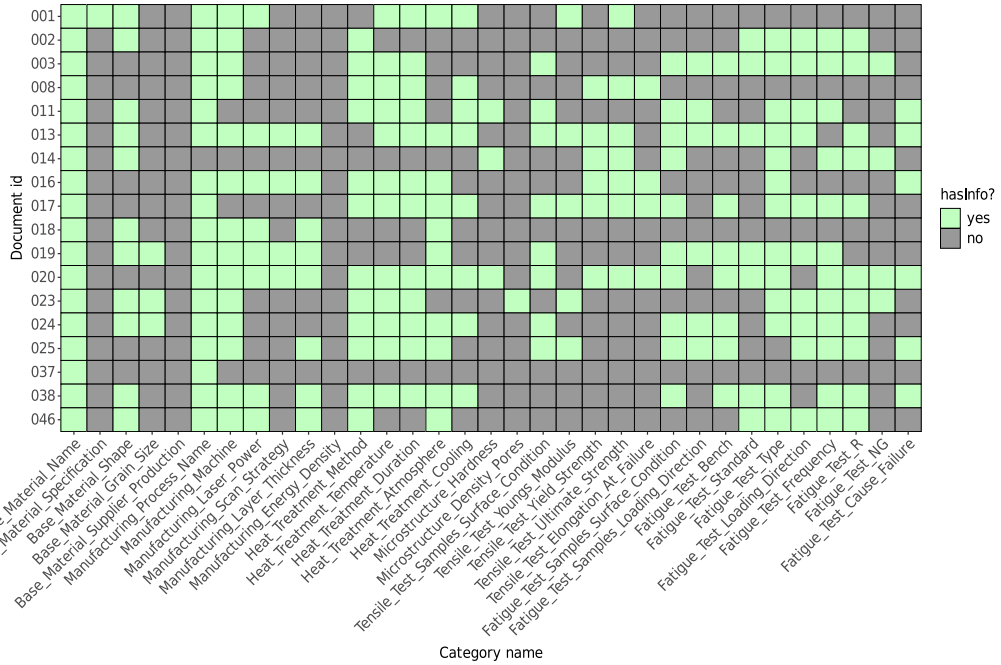

Figure 5. Overview of the detected information for the provided data basis and the chosen search categories using the information extraction method "chatPdf". In the tabular schema, elements are highlighted in green if information was successfully detected, and in gray if no information could be identified.

Table 7. Results of the ChatPDF-based information extraction method "chatPdf" for a given document of the data basis and selected categories.

Id	Search Category Name	Value
1	Name of used for base material	Ti-6Al-4V
2	Specification of base material	String
3	Shape of base material	Tubular
4	Grain size in micrometer of base material	Not specified
5	Name of supplier who produced base material	String
6	Name of manufacturing process	Laser Powder Bed Fusion
7	Name of machine used for manufacturing	Renishaw AM 250
8	Laser power in Watts used for manufacturing	400 W
9	Scan strategy used for manufacturing	Not specified
10	Layer thickness in micrometer used for manufacturing	50 µm
11	Energy density in J/mm^3 used for manufacturing	Not specified
14	Duration in hours of heat treatment	Not specified
15	Atmosphere during heat treatment	Argon
16	Name of cooling type during heat treatment	Not specified
17	Measured Vicker's hardness in HV	Not specified
18	Measured pore density measure during microstructure test	Not specified

Table 7. *Cont.*

Id	Search Category Name	Value
26	Test bench used in fatigue test	Not specified
27	Standard of fatigue test	ASTM Standard
28	Type of fatigue test	Axial-torsion
29	Loading direction in fatigue test	In-phase axial-torsion
30	Test frequency in fatigue test	0.25 to 12 Hz
31	Stress ratio R in fatigue test	R = −1
32	Maximum number of cycles in fatigue test	Not specified
33	Cause of failure of the sample in the fatigue test	Not specified

5. Discussion

This section presents a critical discussion and evaluation of the results obtained from the different information extraction methods. First, Section 5.1 compares the performance of the proposed automatic extraction method with manual extraction by domain experts. Subsequently, the effectiveness of the ChatPDF-based extraction is assessed in relation to manual extraction. These comparisons offer insights into the strengths and weaknesses of each approach. The discussion then addresses the limitations encountered during the study, followed by an analysis of the practical benefits and potential applications of the proposed approach.

5.1. Comparisons

To evaluate the effectiveness of the proposed information extraction method, the information detected is compared against a reference method. The comparison can have the following scenarios:

1. True positive (TP): Both the proposed and the reference method successfully identify the information. This corresponds to cases where the entry is filled in both tables.
2. False positive (FP): The proposed method identifies information (entry filled in the table created using the proposed method), but the reference method does not (entry empty in the table created using the reference method).
3. False negative (FN): The proposed method does not identify information (entry empty in the table created using the proposed method), but the reference method does (entry filled in the table created using the reference method).
4. True negative (TP): Both the proposed and the reference method successfully do not identify the information. This corresponds to cases where the entry is empty in both tables.

As discussed in Section 2, the evaluation of information extraction methods commonly uses precision, recall, and F1 scores. These metrics are derived from the counts of true positives, true negatives, false positives, and false negatives, which are summarized in confusion matrices. Thus, the corresponding confusion matrices are set up, and the performance measures are computed.

5.1.1. Automatic vs. Manual

The results of the automatic information extraction method are compared to the reference method, the manual information extraction method, and are arranged in a tabular schema. The table is provided in Figure 6.

It is clear that some categories of search queries lend themselves to easier information extraction than others. For instance, data relating to material names is usually straightforward to identify. In contrast, extracting values for physical quantities that share the same units or have similar value ranges can be much more complex. A good example of this

is the frequent pairing of yield strength and ultimate tensile strength. These quantities often appear together, have similar names, overlapping value ranges, and use the same units, which can lead to ambiguities and increase the risk of misclassifying the extracted information. The confusion matrix for the automatic information extraction is provided in Figure 7.

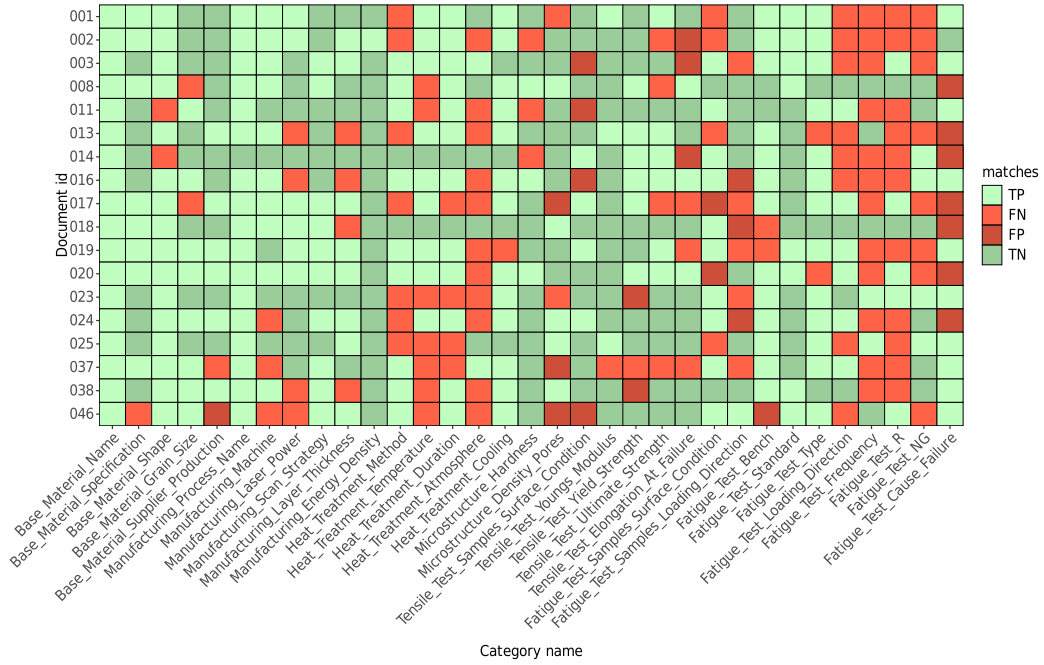

Figure 6. Overview of the detected information for the given dataset and selected search categories, comparing the proposed automatic information extraction method ("automatic") with the reference method ("manual"). Table elements are color-coded according to match type: true positives (TPs) are highlighted in light green, true negatives (TNs) in dark green, false negatives (FNs) in light orange, and false positives (FPs) in dark orange.

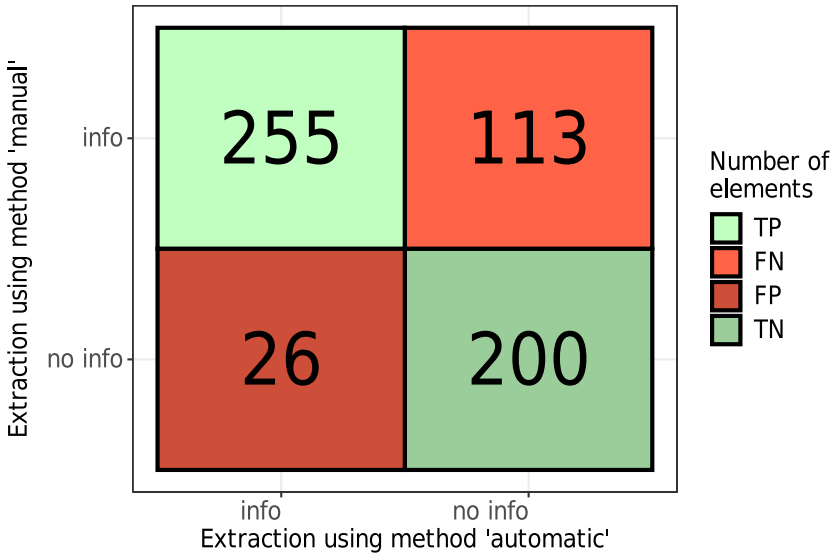

Figure 7. Confusion matrix summarizing the performance of the proposed automatic information extraction method "automatic" in comparison to the manual extraction results. The outcomes are categorized into "no info" and "info" classes, with the matrix detailing the number of elements corresponding to True Positives (TP), False Positives (FP), False Negatives (FN), and True Negatives (TN) for each classification.

When the information extraction algorithm identifies more information than is actually present in the articles (case: false positives), this discrepancy can often be attributed to the overlap between selected search categories, where multiple categories share similar target patterns. False positives most commonly occurred when regular expressions matched contextually irrelevant segments, such as parameter mentions in figure captions or reference lists.

A review of the confusion matrix in Figure 7 reveals a low FP rate of approximately 4.4%. This low rate has several important implications:

Minimal Impact on Data Quality: A 4.4% FP rate indicates that only a small proportion of irrelevant or incorrect information is included in the systematic literature review results.

Limited Additional Manual Effort: The need for researchers to review and filter false positives is reduced compared to higher FP rates, preserving the efficiency gains of the automated workflow.

Lower Risk of Misleading Analyses: The relatively low number of false positives minimizes the potential for inaccuracies in subsequent analyses or meta-studies, supporting the reliability of scientific conclusions.

Continued Room for Optimization: While the workflow already demonstrates strong extraction performance, further reducing the FP rate could enhance the system's trustworthiness and practical value even more.

Conversely, if the algorithm detects less information than is present in the articles (case: false negatives), this may result from poorly chosen target patterns for the search categories or the presence of artefacts such as typographical errors within the articles. This typically occurs in situations where the data appears in atypical phrasings, non-standard formats, or is distributed across multiple sentences—challenges that rule-based methods often cannot adequately address.

The presence of a substantial number of false negatives, as shown in Figure 7, represents a significant limitation, particularly in systematic reviews, where the completeness of data inclusion is essential. An observed false negative (FN) rate of approximately 19% means that a considerable portion of relevant information is not being captured by the extraction process. This has several important implications:

Noticeable Information Loss: A 19% FN rate indicates that a substantial amount of pertinent data are omitted, which can negatively impact the overall quality of the extracted dataset.

Impact on Completeness: Such a high rate of missed values may result in significant gaps, thereby compromising the comprehensiveness and reliability of systematic reviews or meta-analyses. Inadvertent exclusion of relevant studies or critical information could introduce bias into the review outcomes.

Clear Need for Improvement: These findings highlight the inherent limitations of rule-based extraction approaches and underscore the need for integrating more advanced natural language processing techniques to reduce missed values and improve accuracy.

To mitigate these limitations, it is advisable for users to include a manual verification step, especially for cases flagged as low-confidence or excluded by the automated workflow. This additional check can help ensure that critical information is not overlooked, thereby supporting the integrity and reliability of a systematic review process. Table 8 provides a comprehensive overview of the performance of the proposed automatic information extraction method. The accuracy of 0.766 indicates that around 77% of all predictions, whether positive or negative, are correct. Precision is particularly high at 0.908, meaning that the majority of extracted data points identified as relevant are indeed correct, which

is crucial for minimizing the inclusion of irrelevant information (false positives). Recall, at 0.693, shows that approximately 69% of all relevant information present in the articles is successfully captured. This suggests there is still a noticeable proportion of relevant data that the algorithm fails to retrieve, as discussed in the context of false negatives. Specificity (0.885) demonstrates that the method is also effective at correctly identifying irrelevant information, further reducing the risk of false positives. The F1 measure, which balances precision and recall, is 0.786. This value underscores the method's overall ability to maintain a good trade-off between finding as much relevant information as possible while keeping the rate of incorrect extractions low. Overall, these metrics show that the method is well-suited for supporting systematic literature reviews, offering a high level of precision and specificity, and a solid balance between recall and precision. However, there remains room for improvement, particularly in increasing recall to ensure even more comprehensive data extraction.

Table 8. Performance measures related to the evaluation of the proposed information extraction method "automatic".

Measure	Definition/Formula	Value
Accuracy A	(TP + TN)/(TP + FN + FP + TN)	0.766
Precision P	TP/(TP + FP)	0.908
Recall R	TP/(TP + FN)	0.693
Specificity S	TN/(TN + FP)	0.885
F1 measure	2 × (P × R)/(P + R)	0.786

5.1.2. ChatPdf vs. Manual

The results of the ChatPDF-based information extraction method are compared to the reference method, the manual information extraction method, and are arranged in a tabular schema. The table is provided in Figure 8. In a manner analogous to automatic information extraction, it has been observed that the ease of extracting or detecting information varies across different search categories. For instance, data pertaining to material names is generally more accessible. In contrast, information such as the shape, grain size, or supplier of the base material is often more challenging to obtain. Notably, values associated with categories such as the applied frequency, test ratio R, and the maximum number of cycles in fatigue tests can be identified with greater accuracy using the present method than with the proposed automatic information extraction method. Furthermore, the extraction of values for physical quantities that possess identical units and/or similar value domains does not present a significant challenge for the ChatPDF-based extraction method. The confusion matrix for the ChatPDF extraction method is given in Figure 9. In contrast to the proposed rule-based information extraction approach, ChatPDF exhibits a distinct error profile. Specifically, the confusion matrix reveals a false negative (FN) rate of 126 out of 594 cases (approximately 21%) and a very low false positive (FP) rate of 3 out of 594 cases (about 0.5%). The implications of these findings largely mirror those described for the rule-based method: a high FN rate leads to information loss and reduced completeness, which can compromise the comprehensiveness of systematic reviews. However, the exceptionally low FP rate of ChatPDF means that the vast majority of extracted information is relevant, minimizing the need for manual filtering and supporting data quality. Despite this strength, the persistent issue of missed relevant data due to the high FN rate remains. As with the proposed algorithm, further improvements are recommended to reduce false negatives and achieve a more balanced and robust extraction performance. This can include integration of advanced natural language processing techniques. Table 9 provides a concise overview of

key performance metrics for the ChatPDF extraction method, highlighting both its strengths and limitations:

- The very high precision (0.988) and specificity (0.987) indicate that ChatPDF rarely extracts irrelevant information, making its results highly reliable.
- The accuracy (0.783) and F1 score (0.790) are solid, but reflect the imbalance between the extremely high precision and the comparatively lower recall.
- The recall (0.659) is noticeably lower, meaning that a significant proportion of relevant information is missed (a higher rate of false negatives).

The ChatPDF-based extraction method delivers highly precise and clean results, but misses a considerable amount of relevant data. For applications where completeness is essential, it may be advisable to combine this approach with other methods or to include a manual post-processing step.

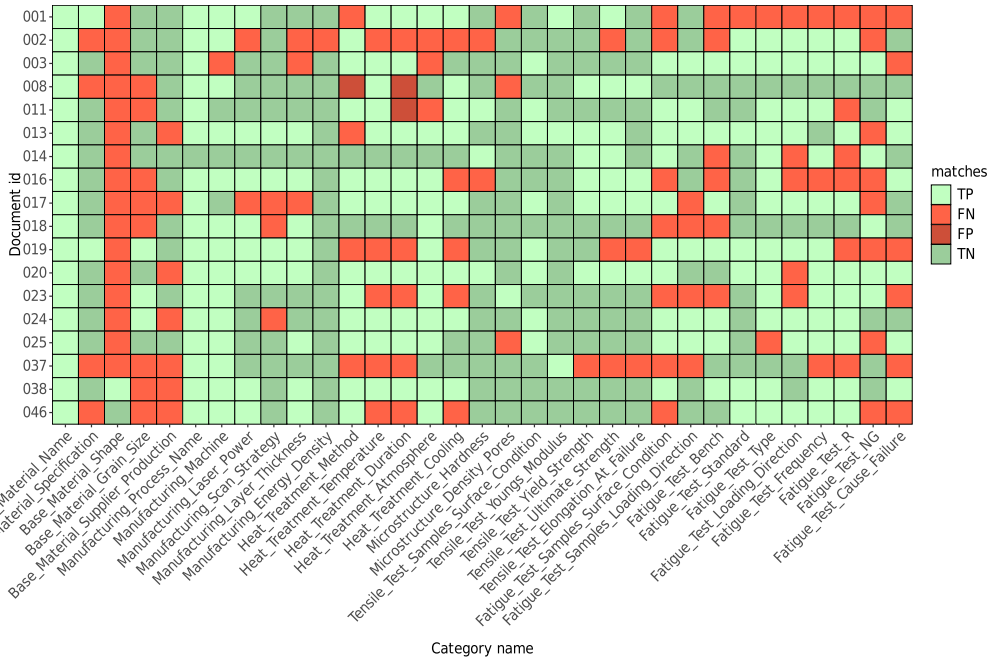

Figure 8. Overview of the detected information for the given dataset and selected search categories, comparing the ChatPDF-based information extraction method ("chatPdf") with the reference method ("manual"). Table elements are color-coded according to match type: true positives (TPs) are highlighted in light green, true negatives (TNs) in dark green, false negatives (FNs) in light orange, and false positives (FPs) in dark orange.

Table 9. Performance measures related to the evaluation of information extraction method "chatPdf".

Measure	Definition/Formula	Value
Accuracy A	(TP + TN)/(TP + FN + FP + TN)	0.783
Precision P	TP/(TP + FP)	0.988
Recall R	TP/(TP + FN)	0.659
Specificity S	TN/(TN + FP)	0.987
F1 measure	2 × (P × R)/(P + R)	0.790

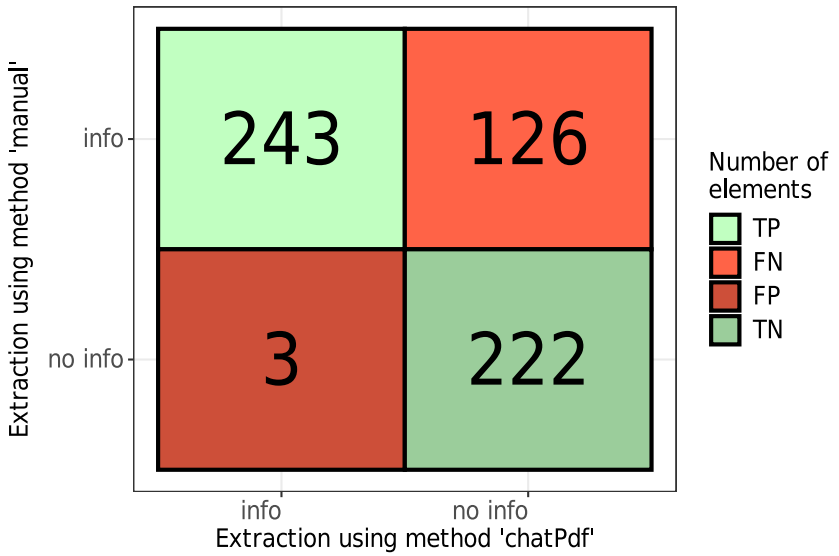

Figure 9. Confusion matrix summarizing the performance of the ChatPDF-based information extraction method "chatPdf" in comparison to the manual extraction results. The outcomes are categorized into "no info" and "info" classes, with the matrix detailing the number of elements corresponding to True Positives (TP), False Positives (FP), False Negatives (FN), and True Negatives (TN) for each classification.

5.2. Benefits

Efficiency: The automatic extraction tool dramatically reduces the time required to process and analyze scientific publications. What previously took hours of manual work can now be completed in seconds per document, enabling researchers to focus on higher-level analysis and synthesis.

Scalability: The workflow is capable of handling large datasets, making it suitable for systematic reviews or meta-analyses involving hundreds of publications. This scalability facilitates comprehensive literature coverage and supports data-driven research.

Accessibility: Designed for use on standard hardware and without the need for advanced programming skills, the tool lowers the barrier for adoption among researchers from various backgrounds.

Consistency and Reproducibility: Automated extraction ensures consistent application of search patterns and rules, reducing human error and subjective bias. This enhances the reproducibility of literature reviews and data extraction processes.

Higher Recall and Flexibility: The proposed rule-based approach achieves a higher recall than the ChatPDF-based method, meaning it successfully captures a greater proportion of relevant information present in the articles. This reduces the risk of omitting important data and is particularly advantageous for systematic reviews where completeness is critical. The method is well-suited for scenarios where sensitivity and comprehensive data collection are prioritized over absolute precision, ensuring that fewer relevant studies or data points are missed.

Customizability: Through the use of regular expressions and configurable search patterns, the tool can be adapted to different research domains and evolving information needs.

Open Science Potential: The planned open-source release and possible web-based implementation foster collaboration, transparency, and continuous improvement within the research community.

Data Security and Local Processing: The proposed approach does not require specialized hardware and can be executed on a standard local client. Local processing ensures

a high level of data security, as sensitive documents do not need to be transferred to third parties. Additionally, the method enables rapid and efficient processing of typical scientific publications.

User-Friendliness: The intuitive interface and straightforward workflow allow researchers to manage and organize large document collections efficiently, even without technical expertise.

5.3. Limitations

Data Quality Dependency: The accuracy of extraction is highly dependent on the quality of the text extracted from PDFs. Poor OCR results, corrupted files, or inconsistent PDF structures can lead to incomplete or erroneous data.

Complex Document Layouts: Multi-column formats, embedded figures, tables, and non-standard layouts present significant challenges for accurate block detection and semantic preservation. This may result in fragmented or misaligned information extraction.

Lower Precision and Manual Effort Required: Compared to ChatPDF, the rule-based method demonstrates lower precision and specificity, which leads to a higher rate of false positives. This means that more irrelevant or incorrectly classified information is included in the results. The inclusion of irrelevant data requires additional manual review and filtering by researchers, which can partially negate the efficiency gains achieved through automation. While recall is higher than with ChatPDF, the overall balance between precision and recall (as reflected in the F1 score) may be less favorable, indicating a trade-off between completeness and the need for manual post-processing.

Effort Required for Defining Search Patterns: Implementing the approach necessitates a certain degree of effort, particularly in defining and adjusting suitable regular expressions ("keys"). Developing regular expressions for physical quantities involving floating-point numbers and units is especially complex. While the selection of categories is generally appropriate, distinguishing between categories with similar target entities can be challenging.

Search Pattern Sensitivity: The effectiveness of extraction relies heavily on the definition and specificity of search patterns. Inflexible or poorly defined patterns may miss relevant information or generate false positives.

Limited Contextual Understanding: The tool currently lacks sophisticated mechanisms for contextual analysis, which can lead to misinterpretation of ambiguous or domain-specific terms. Human review is still necessary to ensure the relevance and accuracy of extracted data.

Manual Pre- and Post-Processing: Some manual intervention may still be required for pre-processing (e.g., handling protected PDFs, ensuring file quality) and post-processing (e.g., validation, data cleaning), which can limit the overall automation benefit.

Domain Adaptation: While the tool is customizable, significant adaptation effort may be needed to apply it to domains with very different terminology, reporting standards, or document structures.

Handling of Non-Textual Data: Extraction of information from images, complex tables, or graphical elements remains limited, potentially omitting valuable data present in figures.

Legal and Ethical Considerations: Automated extraction from copyrighted or paywalled documents may raise legal or ethical issues, especially when sharing or publishing extracted data.

Scalability Constraints: Although scalable for moderate datasets, extremely large-scale deployments may require additional optimization or parallel processing capabilities.

5.4. Ethical and Copyright Implications

The automated extraction of information from scientific publications raises several ethical and legal considerations. Firstly, copyright law in many jurisdictions protects the layout, structure, and content of scientific articles, even when they are openly accessible [34]. Automated extraction methods that process full texts may risk infringing these rights, especially when applied to subscription-based or embargoed content. Researchers must ensure that their data sources either permit such uses under their licensing terms (e.g., via Creative Commons licenses) or that explicit permission has been obtained [35].

From an ethical perspective, the mass extraction and aggregation of data could potentially undermine the interests of publishers, authors, and journals, particularly if the extracted content is redistributed or used to circumvent paywalls. It is essential to respect the intellectual property and moral rights of authors, and to use extracted data solely for legitimate scholarly purposes.

Academic associations such as the International Association of Scientific, Technical, and Medical Publishers (STM) and the Committee on Publication Ethics (COPE) have issued guidelines emphasizing responsible text and data mining. These include respecting licenses, ensuring proper attribution, and not engaging in systematic downloading that could disrupt publisher services [36,37]. Many associations encourage open access and support responsible data mining, provided that it does not infringe upon copyright or breach data privacy.

In summary, while automated information extraction offers significant benefits for research efficiency and reproducibility, it is crucial to operate within the bounds of copyright law and ethical guidelines. Researchers should always verify the legal status of their data sources and adhere to the recommendations of relevant academic associations.

6. Conclusions

6.1. Summary

This study aimed to develop and validate an automatic method for extracting information from scientific articles in PDF format, with a particular emphasis on the field of AM. The effectiveness of the approach was demonstrated through its application to real-world documents, highlighting its ability to organize unstructured collections and support the identification of relevant parameters.

The method was evaluated using a dataset of 18 open-access scientific publications from diverse journals and conference proceedings within the AM domain. The results of the automated extraction were benchmarked against both manual extraction and a contemporary LLM-based approach. The findings show that the proposed workflow can accurately and efficiently extract relevant process and material data, offering performance that is competitive with the LLM-based method.

The main research questions posed in the Introduction section have been systematically addressed:

RQ1: Automated Parameter Extraction: A structured methodology was implemented to determine the extent to which parameters—such as physical ranges and material properties—can be automatically extracted from scientific publications. Tailored search patterns and algorithms were developed and applied, as described in Section 3.2.3. The outcomes, including the efficiency and accuracy of parameter extraction, are presented and discussed in Sections 4 and 5.

RQ2: Reliability of Extracted Information: The reliability of the automatically extracted information was evaluated by comparison with manually curated data. This assessment, detailed in Section 5, provided insights into the system's precision and recall, highlighting its strengths and areas for improvement.

RQ3: Category-Specific Information Extraction: The feasibility of extracting information for predefined categories was explored by designing and applying category-specific search patterns, as outlined in Section 3.1.3. The study demonstrated that, given well-defined search criteria, the system can effectively extract targeted information across multiple categories.

In summary, it was demonstrated that the proposed approach can deliver fast and accurate results on a local PC, without the need for specialized hardware. This makes it a practical and accessible solution for researchers and practitioners alike.

6.2. Key Benefits

The proposed approach enables the rapid and accurate extraction of relevant information from scientific PDF documents, greatly reducing the manual effort and time required. Key benefits of the tool are as follows:

Efficiency: Automated extraction reduces processing time from hours to seconds per document, allowing researchers to quickly process large volumes of literature.
Accessibility: The tool is designed for use on standard hardware and does not require advanced technical expertise, making it accessible to a broad user base.
Accuracy: The method delivers a high match rate with manual extraction, ensuring reliable and precise results.
Scalability: The workflow is suitable for large document collections and can be adapted for use in other scientific domains beyond additive manufacturing.
Open Science: With a planned open-source release and potential web-based implementation, the tool fosters collaboration and broad accessibility within the research community.
User-Friendliness: Researchers can import PDF files in a straightforward and intuitive manner, making the tool highly effective for managing and organizing existing collections of documents.

6.3. Key Limitations

The study identified several key limitations that currently affect the performance of the approach:

Data Quality: The reliability of the extraction process is highly dependent on the quality of the text obtained from PDF documents. Errors in recognizing or segmenting PDF elements can disrupt semantic coherence, making high-quality input data essential for accurate results.
Definition of Search Patterns: The effectiveness of information extraction is strongly influenced by the precision of the search patterns used. Poorly defined or overly broad patterns may result in missed or irrelevant information.
Document Layout: Variations in document formatting—such as multi-column layouts or inconsistent journal structures—present significant challenges for accurate extraction. These layout differences can hinder the correct preservation of text semantics. Enhancing data pre-processing and exploring AI-based methods for block type identification could improve extraction quality.
Contextual Relevance: While the tool provides a robust foundation for identifying target information, it does not yet reliably determine the contextual relevance of extracted information. At this stage, human evaluation remains necessary to ensure the applicability and accuracy of the results.

6.4. Future Work and Outlook

A comprehensive evaluation of the validity and generalisability of the proposed approach requires its application to scientific publications in domains beyond additive manufacturing. The algorithm's design, enabling reconfiguration of search categories and the use of alternative regular expressions, supports the expectation of methodological transferability. Nonetheless, empirical validation and optimization in a range of scientific fields are still needed to confirm the approach's effectiveness and robustness.

While comprehensive literature reviews sometimes yield inconclusive results, they can also highlight gaps unaddressed by automatic data extraction. Investigating whether quantitative and generic automatic data extraction enhances or diminishes the quality of data obtained by researchers will be crucial. This analysis could provide insights into the balance between automation and manual intervention in data extraction processes.

Developing a fully automatic information extraction system could eliminate the need for human evaluation. Achieving this goal requires mapping extracted information to domain-specific ontologies or taxonomies, supported by controlled vocabularies and synonym dictionaries. This would enable the transformation of sentences into structured triples (e.g., RDF triples) for comparison against a knowledge base. Such advancements would result in a versatile tool capable of evaluating extensive text-based information, empowering users to enhance their knowledge base efficiently. To ensure the system's utility across diverse domains, it must be standardized and readily expandable. This would facilitate the integration of extensive datasets and ensure the system's adaptability to a wide range of use cases, fostering its role as a multi-purpose tool for data analysis.

Implementing a feedback reporting system would further support continuous improvement by allowing users to identify and correct extraction errors.

The outsourcing of information extraction processes prompts a critical question: how might the substitution of these processes impact the knowledge base of the research community? Automatic information extraction, as previously discussed, facilitates the bypassing of certain manual processes, thereby enhancing specific aspects of research methodologies and allowing researchers to focus on primary parameters. However, this shift may inadvertently lead to the exclusion of information that previously played a significant role in shaping research processes. This potential issue necessitates rigorous examination in future analyses of scientific publications. Such investigations should aim to establish an informational foundation to determine whether automatic information extraction should be governed by defined limitations and boundaries, and to assess whether the advantages of automation outweigh its potential drawbacks.

To mitigate these challenges, it is essential to develop and implement standards for information extraction. This should include both manual and automatic validation and verification procedures to ensure the quality and completeness of the extracted information. These measures would not only help safeguard the integrity of the research community's knowledge base but also provide a structured framework for applying automatic extraction methods effectively across various domains and use cases.

The process of extracting information from PDF documents becomes significantly more straightforward when the structure of the documents is known, compared to cases where the structure is undefined or inconsistent. A potential solution to enhance the extraction of data from scientific works in PDF format is the standardization of content structure. Generally, the structure of scientific works on a specific subject exhibits a degree of consistency. To achieve this improvement, it would be necessary for publishers to establish and adopt a unified standard for structuring articles published in specialist journals and conference proceedings.

Furthermore, the source code for the proposed solution method is planned to be made available to the research community under an open-source license. This will encourage collaboration and further development. Additionally, implementing the solution method as a web-based application would significantly simplify and systematize the process of conducting literature research, providing researchers with an efficient and accessible tool for information extraction.

The integration of the proposed automated information extraction method into systematic review workflows holds significant promise for streamlining the synthesis of scientific evidence. Although the algorithm is highly effective at extracting data from selected publications, future enhancements could aim to support earlier stages of the review process, such as automated screening and pre-selection of studies according to predefined inclusion and exclusion criteria. By combining advanced natural language processing techniques with transparent documentation and registration practices, as mandated by systematic review protocols, both the efficiency and reproducibility of evidence synthesis across a range of scientific fields may be further improved.

Author Contributions: Conceptualization, K.F., H.W. and R.K.; Data curation, K.F., R.K. and P.T.; Formal analysis, K.F., R.K. and P.T.; Funding acquisition, H.W., R.K., S.I. and M.Z.; Investigation, K.F. and P.T.; Methodology, K.F., H.W., R.K. and P.T.; Project administration, K.F., R.K. and P.T.; Resources, K.F., R.K. and P.T.; Software, K.F.; Supervision, S.I. and M.Z.; Validation, K.F., R.K. and P.T.; Visualization, K.F.; Writing —— original draft, K.F., R.K. and P.T.; Writing —— review and editing, K.F., H.W., R.K., P.T., S.I. and M.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research has been funded in parts within the research projects "AMTwin" and "Werk4.0". "AMTwin" (Data-driven process, material, and structure analysis for additive manufacturing) has been funded by the Sächsische Aufbaubank (SAB) via funds of the European Regional Development Fund (ERDF) and co-financed with tax revenue based on the budget approved by the parliament of the Free State of Saxony (Germany), grant numbers 100373334, 100373343. "Werk4.0" has been funded by the German Federal Ministry of Economics and Climate Protection (BMWK) in the funding guideline "Digitalization of Vehicle Manufacturers and Supplier Industry" in the funding framework "Future Investments Vehicle Manufacturers and Supplier Industry". It is supervised by the project sponsor VDI Technologiezentrum GmbH (grant number 13IK022K).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset comprises 18 scientific publications related to Ti6Al4V, in PDF format, published between 2015 and 2018. All articles are accessible online and are designated as open access. The names of the journals or conferences are not provided in their original form; instead, they are represented by single letters. Moreover, the names of the authors are absent from the dataset. The anonymization of the dataset was conducted with the objective of focusing on the fundamental aspects of data quality, without the potential influence of author or journal/conference names. Accordingly, the data are not publicly accessible; they are only available upon request from the corresponding author.

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

References

1. Chowdhury, G.G. *Introduction to Modern Information Retrieval*; Facet Publishing: London, UK, 2010.
2. Nasar, Z.; Jaffry, S.W.; Malik, M. Information extraction from scientific articles: A survey. *Scientometrics* **2018**, *117*, 1931–1990. [CrossRef]
3. Zhu, R.; Tu, X.; Xiangji Huang, J. Chapter seven—Deep learning on information retrieval and its applications. In *Deep Learning for Data Analytics*; Das, H., Pradhan, C., Dey, N., Eds.; Academic Press: Cambridge, MA, USA, 2020; pp. 125–153. [CrossRef]
4. Esteva, A.; Kale, A.; Paulus, R.; Hashimoto, K.; Yin, W.; Radev, D.; Socher, R. COVID-19 information retrieval with deep-learning based semantic search, question answering, and abstractive summarization. *npj Digit. Med.* **2021**, *4*, 68. [CrossRef] [PubMed]

5. Grace, S.; Rosenthal, J. *Sourcing and Referencing*; Brill: Leiden, The Netherlands, 2009. [CrossRef]
6. Chen, C. Science Mapping: A Systematic Review of the Literature. *J. Data Inf. Sci.* **2017**, *2*, 1–40. [CrossRef]
7. Gusenbauer, M. Google Scholar to Overshadow Them All? Comparing the Sizes of 12 Academic Search Engines and Bibliographic Databases. *Scientometrics* **2019**, *118*, 177–214. [CrossRef]
8. Marcos-Pablos, S.; García-Peñalvo, F. Information retrieval methodology for aiding scientific database search. *Soft Comput.* **2020**, *24*, 5551–5560. [CrossRef]
9. Wang, Y.; Wang, L.; Rastegar-Mojarad, M.; Moon, S.; Shen, F.; Afzal, N.; Liu, S.; Zeng, Y.; Mehrabi, S.; Sohn, S.; et al. Clinical Information Extraction Applications: A Literature Review. *J. Biomed. Inform.* **2017**, *77*, 34–49. [CrossRef] [PubMed]
10. Welcome to Python.org. Available online: https://www.python.org/ (accessed on 29 June 2025).
11. R: The R Project for Statistical Computing. Available online: https://www.r-project.org/ (accessed on 29 June 2025).
12. Tkaczyk, D.; Szostek, P.; Fedoryszak, M.; Dendek, P.J.; Bolikowski, L. CERMINE: Automatic extraction of structured metadata from scientific literature. *Int. J. Doc. Anal. Recognit. (IJDAR)* **2015**, *18*, 317–335. [CrossRef]
13. Ahmed, I. Astera ReportMiner. Available online: https://www.astera.com/products/report-miner/ (accessed on 29 June 2025).
14. Gwizdka, J.; Hansen, P.; Hauff, C.; He, J.; Kando, N. Search as Learning (SAL) Workshop 2016. In Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, Pisa Italy, 17–21 July 2016; SIGIR '16, pp. 1249–1250. [CrossRef]
15. Müller, H.M.; Kenny, E.; Sternberg, P. Textpresso: An Ontology-Based Information Retrieval and Extraction System for Biological Literature. *PLoS Biol.* **2004**, *2*, e309. [CrossRef] [PubMed]
16. Ceci, F.; Pietrobon, R.; Goncalves, A. Turning Text into Research Networks: Information Retrieval and Computational Ontologies in the Creation of Scientific Databases. *PLoS ONE* **2012**, *7*, e27499. [CrossRef] [PubMed]
17. Dragoni, M.; da Costa Pereira, C.; Tettamanzi, A. A Conceptual Representation of Documents and Queries for Information Retrieval Systems by Using Light Ontologies. *Expert Syst. Appl.* **2012**, *39*, 10376–10388. [CrossRef]
18. Lopez, P. GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In Proceedings of the 13th European Conference on Research and Advanced Technology for Digital Libraries, Corfu, Greece, 27 September–2 October 2009. [CrossRef]
19. Welcome to Pdfminer.six's Documentation!—Pdfminer.six__VERSION__Documentation. Available online: https://pdfminersix.readthedocs.io/en/latest/ (accessed on 29 June 2025).
20. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the NAACL-HLT, Minneapolis, MN, USA, 2–7 June 2019. [CrossRef]
21. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), Long Beach, CA, USA, 4–9 December 2017; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp. 6000–6010. [CrossRef]
22. Kardas, M.; Borchmann, L.; Bembenek, P.; Lewandowski, M.; Marcinczuk, M.; Gawor, M.; Rybak, P.; Wroblewska, A.; Rychlikowski, P.; Kocon, J. Document Structure Recognition: A Review. *arXiv* **2020**, arXiv:2008.05961. [CrossRef]
23. Ponte, J.; Croft, W. A Language Modeling Approach to Information Retrieval. *ACM SIGIR Forum* **2017**, *51*, 202–208. [CrossRef]
24. Zhang, W.; Zhao, X.; Zhao, L.; Yin, D.; Yang, G.H.; Beutel, A. Deep Reinforcement Learning for Information Retrieval: Fundamentals and Advances. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event China, 25–30 July 2020; SIGIR '20, pp. 2468–2471. [CrossRef]
25. Tshitoyan, V.; Dagdelen, J.; Weston, L.; Dunn, A.; Rong, Z.; Kononova, O.; Persson, K.A.; Ceder, G.; Jain, A. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **2019**, *571*, 95–98. [CrossRef] [PubMed]
26. Authors of chatpdf.com. ChatPDF—Chat with any PDF! Available online: https://www.chatpdf.com/ (accessed on 29 June 2025).
27. Kovacevic, A.; Ivanovic, D.; Milosavljević, B.; Konjovic, Z.; Surla, D. Automatic extraction of metadata from scientific publications for CRIS systems. *Program Electron. Libr. Inf. Syst.* **2011**, *45*, 376–396. [CrossRef]
28. Dieb, S.; Yoshioka, M.; Hara, S.; Newton, M. Framework for automatic information extraction from research papers on nanocrystal devices. *Beilstein J. Nanotechnol.* **2015**, *6*, 1872–1882. [CrossRef] [PubMed]
29. ScienceBeam: Using Open Technology to Extract Knowledge from Research PDFs. Available online: https://elifesciences.org/labs/743da0fc/sciencebeam-using-open-technology-to-extract-knowledge-from-research-pdfs (accessed on 29 June 2025).
30. Raßloff, A.; Feldhoff, K.; Wiemer, H.; Zimmermann, M.; Kästner, M. AMTwin-Datengetriebene Prozess-, Werkstoff- und Strukturanalyse für die additive Fertigung. In Proceedings of the Mobilität der Zukunft–Bauteilzuverlässigkeit im digitalen Zeitalter-DVM-Tag 2023, Berlin, Germany 29–30 March 2023. [CrossRef]
31. Li, P.; Warner, D.H.; Fatemi, A.; Phan, N. Critical assessment of the fatigue performance of additively manufactured Ti–6Al–4V and perspective for future research. *Int. J. Fatigue* **2016**, *85*, 130–143. [CrossRef]
32. Liu, S.; Shin, Y.C. Additive manufacturing of Ti6Al4V alloy: A review. *Mater. Des.* **2019**, *164*, 107552. [CrossRef]
33. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2023.

34. Crews, K.D. Copyright and the academic author: Legal rights and responsibilities. *Portal Libr. Acad.* **2012**, *12*, 297–317.
35. Creative Commons. About The Licenses. Available online: https://creativecommons.org/licenses/ (accessed on 6 August 2025).
36. International Association of Scientific, Technical and Medical Publishers. STM Text and Data Mining Guidelines. Available online: https://www.stm-assoc.org/intellectual-property/text-and-data-mining/ (accessed on 6 August 2025).
37. Committee on Publication Ethics. COPE Guidelines. Available online: https://publicationethics.org/guidance/Guidelines (accessed on 6 August 2025).

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