MESD: Metadata Extraction from Scholarly Documents – A Shared Task Overview

Zeyd Boukhers[1,2,*], Cong Yang[3]

[1]Fraunhofer Institute for Applied Information Technology, Sankt Augustin, Germany
[2]University Hospital of Cologne, Cologne, Germany
[3]Soochow University, China

Abstract

This paper presents an overview of the Metadata Extraction from Scholarly Documents (MESD) shared task, which was designed to address the challenge of extracting structured metadata (e.g. Title, Author, Abstract, etc.) from scientific publications. The task aimed to promote the development of techniques for making scholarly data more Findable, Accessible, Interoperable, Reusable (FAIR) by improving metadata extraction from PDF documents. We describe the task design and the creation of two complementary datasets: (1) the S2ORC_Exp500v1 dataset consisting of 500 training samples, 100 validation samples, and 100 test samples with text-based annotations, and (2) the SSOAR Multidisciplinary Vision Dataset (SSOARGMVD) containing more than 8000 documents with bounding box annotations suitable for computer vision approaches. We discuss potential directions for future research in metadata extraction from scholarly documents, highlighting the opportunities presented by these new resources.

Keywords

metadata extraction, document processing, scholarly documents, natural language processing

1. Introduction

Scientific literature continues to grow at an unprecedented rate, with millions of new scholarly documents published each year. Making this vast corpus of knowledge discoverable and reusable depends critically on the availability of high-quality metadata. While contemporary publications typically include structured metadata, a significant proportion of the existing scholarly record, particularly from smaller publishers and historical archives, lacks accessible metadata [1, 2]. This gap represents a substantial challenge for scientific information retrieval and knowledge management.

The scientific community has recognized the importance of making research outputs findable, accessible, interoperable, and reusable (FAIR) [3]. Metadata plays a crucial role in achieving these FAIR principles, serving as the foundation for discovery systems, citation networks, and knowledge graphs. Despite this critical role, many scholarly documents remain difficult to discover and integrate into the scientific knowledge ecosystem due to metadata deficiencies.

The Metadata Extraction from Scholarly Documents (MESD) shared task was conceived to address this challenge by encouraging the development of automated techniques for extracting key metadata elements from scientific publications. The primary goal was to advance the state of the art in metadata extraction from PDFs, with a focus on practical applications that could help make scholarly literature more FAIR. The task specifically targeted publications from smaller and mid-sized publishers, which often exhibit greater variability in formatting and layout compared to publications from major publishers with standardized templates.

The MESD shared task focused on extracting key bibliographic metadata elements: title, authors, abstract, keywords, Digital Object Identifier (DOI), publication venue, publication date, volume/issue

2nd International Workshop on Natural Scientific Language Processing and Research Knowledge Graphs (NSLP 2025), co-located with ESWC 2025, June 01–02, 2025, Portorož, Slovenia

*Corresponding author.
zeyd.boukhers@fit.fraunhofer.de (Z. Boukhers); cong.yang@suda.edu.cn (C. Yang)
https://zeyd.boukhers.com (Z. Boukhers)
0000-0001-9778-9164 (Z. Boukhers); 0000-0002-8314-0935 (C. Yang)

information, and page numbers. These elements form the foundation of discoverability and citation networks in scholarly communication systems.

In this paper, we describe the motivation, design, and execution of the MESD shared task, including the creation of two specialized datasets and our evaluation methodology. The first dataset, S2ORC_Exp500v1, consists of 500 training samples, 100 validation samples, and 100 test samples with text-based annotations derived from the Semantic Scholar Open Research Corpus. The second dataset, SSOAR Multidisciplinary Vision Dataset (SSOAR-MVD), contains 50,000 documents with bounding box annotations suitable for computer vision approaches, specifically targeting documents from disciplines known for challenging layouts such as Social Sciences, Humanities, Law, and Administration.

We outline the challenges inherent in metadata extraction from scholarly documents, discuss evaluation considerations, and propose directions for future research, including the creation of FAIR Digital Objects based on extracted metadata. By providing these resources and insights, we aim to stimulate further innovation in making scholarly literature more discoverable and reusable.

The remainder of this paper is organized as follows: Section 2 reviews related work in metadata extraction and discusses the challenges in metadata extraction. Section 3 describes the MESD task in detail, including the specific metadata elements targeted. Section 4 outlines the evaluation methodology, while Section 6 proposes future directions, and Section 7 concludes the paper.

2. Related Work

2.1. Metadata Extraction Techniques and Approaches

Metadata extraction from scholarly documents has been an active area of research for over two decades, with approaches evolving from rule-based systems to sophisticated machine learning techniques [4]. This section reviews key developments in the field

Early efforts in metadata extraction relied primarily on handcrafted rules and templates. Systems like ParsCit [5] and GROBID [6] initially employed rule-based strategies to identify metadata from scholarly documents. These approaches performed reasonably well for documents with consistent formatting but struggled with the diversity of layouts and styles found across different publishers and disciplines [7, 8]. Despite these limitations, rule-based components continue to play a role in modern hybrid systems, particularly for well-structured metadata elements like DOIs and dates [9, 10].

The limitations of rule-based systems led to the adoption of machine learning techniques for metadata extraction. Conditional Random Fields (CRFs) became particularly popular for sequence labeling tasks in document processing [11, 12]. CERMINE [9] employed a CRF-based approach for extracting metadata from scientific literature, while Peng and McCallum [11] used CRFs for bibliographic reference extraction. These statistical approaches offered better generalization to unseen document formats compared to rule-based methods [7].

With the advancement of deep learning, neural networks have increasingly been applied to metadata extraction tasks. Bi-directional Long Short-Term Memory (BiLSTM) networks and their variants have shown promising results [13]. For instance, An et al. [14] introduced a DNN-based segment sequence labeling approach that outperformed traditional methods on standard datasets. Chiu and Nichols [15] applied BiLSTM-CNN models for named entity recognition, a technique that has since been adapted for metadata extraction tasks. These deep learning approaches benefit from their ability to learn complex patterns without extensive feature engineering [14, 16].

Recognizing the importance of document layout and visual cues, some researchers have explored computer vision techniques for metadata extraction. DeepPDF [17] pioneered the use of neural networks for PDF document segmentation, treating the task as an image segmentation problem. More recently, approaches like PubLayNet [18] have utilized mask region-based convolutional neural networks (Mask R-CNN) to identify different components of scientific documents based on their visual appearance.

MexPub [19] specifically addressed the challenges of extracting metadata from German scientific publications by using Mask R-CNN to analyze document images. Ali et al. [20] explored computer vision and machine learning approaches for metadata enrichment of historical newspaper collections.

These vision-based approaches have proven particularly effective for documents with complex layouts or when text extraction is unreliable [21, 22].

The most recent trend in metadata extraction involves multimodal approaches that combine textual and visual features. Liu et al. [23] proposed a deep learning architecture that processes both the textual content and the visual layout of documents. Similarly, Boukhers and Bouabdallah [2] introduced a multimodal approach that simultaneously analyzes PDF documents as text and as images.

Balasubramanian et al. [24] demonstrated the superiority of multimodal approaches over unimodal ones in metadata extraction from video lectures, achieving significant improvements in precision and recall. These multimodal approaches represent the current state of the art, as they leverage complementary signals from both the textual content and visual layout of documents. However, they often require larger datasets and more computational resources compared to unimodal approaches [2, 23].

Despite the importance of metadata extraction, there have been relatively few shared tasks specifically focused on this challenge. The 1st Workshop on Scholarly Document Processing included shared tasks on citation contextualization and style [25], but focused primarily on citation contexts rather than document metadata extraction. the BiblioDAP workshop (The 1st Workshop on Bibliographic Data Analysis and Processing) [26] addressed challenges in bibliographic data processing, including metadata extraction and management, though it has a broader scope that includes citation analysis and bibliometric studies as well.

Datasets for metadata extraction have also been limited in size and scope. The UMass citation dataset [27] and Cora reference string dataset [28] have been commonly used for evaluating citation extraction systems, but they focus on bibliographic references rather than document metadata. The PubLayNet dataset [18] provides document layout annotations but does not include specific metadata annotations. The CiteSeerX dataset [29] offers a larger collection but with varying annotation quality [7].

The S2ORC corpus [30] represents a significant advancement, providing a large dataset of open access papers with associated metadata, though it was not specifically designed for training metadata extraction systems. The GROBID dataset [6] includes annotated documents for training but is primarily focused on bibliographic references.

The MESD shared task addresses these gaps by providing two complementary datasets—S2ORC_Exp500v1 with detailed text annotations and SSOAR-MVD with computer vision-oriented bounding box annotations—specifically designed for metadata extraction from scholarly documents. By encompassing both textual and visual approaches, these datasets enable more comprehensive evaluation of metadata extraction techniques.

2.2. Challenges in Metadata Extraction

Metadata extraction from scholarly documents presents several significant challenges that make it an interesting and complex research problem:

Task Complexity and Resource Requirements Metadata extraction from PDFs is an inherently complex task that sits at the intersection of document layout analysis, information extraction, and natural language processing. Developing effective solutions requires expertise in multiple domains and potentially significant computational resources for training models on document images or structured representations. This complexity is particularly evident when processing documents with non-standard layouts [31, 7] or from multiple publishers with different formatting conventions [1, 4].

Variability in Document Formats Scientific publications exhibit considerable variability in their formatting and organization. Journal articles, conference papers, preprints, and technical reports each follow different conventions for presenting metadata. Even within a single document type, there can be substantial variation across publishers, disciplines, and time periods. This heterogeneity necessitates robust approaches that can adapt to different document structures [9, 2, 32].

Data Representation Challenges The transition from visual PDF representation to structured metadata involves multiple transformations, each introducing potential errors. Text extraction from PDFs can suffer from issues like incorrect character recognition, disrupted reading order, and loss of formatting cues that might indicate metadata boundaries. These challenges are compounded when dealing with multi-column layouts [18], embedded figures and tables, or documents with complex mathematical notation.

Evaluation Considerations The evaluation based on Levenshtein Similarity with a 95% threshold represents a stringent standard that acknowledges the importance of accuracy in metadata extraction while allowing for minor variations in text representation. This approach balances the need for precise extraction with the practical realities of processing diverse document formats.

3. MESD Task

The MESD shared task focused on the extraction of nine predefined metadata elements from scholarly documents:

- **Title**: The main title of the publication, including separations such as colons or dashes.
- **Authors**: All named contributors, typically appearing after the title. Multiple authors are separated by commas or "and".
- **Abstract**: The summary paragraph(s) typically appearing at the beginning of the paper, often preceded by an "Abstract" header. When abstract content overlaps with keywords, we consider only the text preceding any keyword listing.
- **Keywords**: Terms indicating the paper's subject matter, typically appearing as a list after the abstract, often preceded by "Keywords:" or similar indicators.
- **DOI**: The persistent identifier in the format "10.xxxx/xxxxx" or "https://doi.org/10.xxxx/xxxxx".
- **Publication venue**: The journal name, conference proceedings, or book title where the paper appeared. We consider the primary publication container (e.g., "Journal of Informatics") as the venue, while conference proceedings details are included in volume/issue information.
- **Publication date**: Any date information related to publication, including online availability, print dates, or submission dates.
- **Volume/issue information**: Numeric or alphanumeric identifiers for the specific volume or issue (e.g., "Vol. 42, No. 3").
- **Page numbers**: The range or single page indicating the paper's location within a larger publication (e.g., "pp. 123-145").

Participants were challenged to develop systems capable of identifying and extracting these elements from the PDF documents. A key aspect of the task was handling the variability in document structures and formats, as well as dealing with potential missing elements (e.g., some documents might not include a DOI or explicit keywords). The task was designed to simulate real-world scenarios where metadata extraction systems must process documents from different publishers, domains, and time periods. The evaluation was based on the system's ability to correctly identify the presence or absence of each metadata element and accurately extract the text content when present.

3.1. Dataset Creation

Two datasets were created specifically for the MESD shared task:

3.1.1. S2ORC_Exp500v1 Dataset

The *S2ORC_Exp500v1* dataset was created specifically for the MESD shared task. The source documents were selected from the S2ORC (Semantic Scholar Open Research Corpus) [30], which provides a large

collection of open-access scientific papers. The selection process aimed to ensure diversity in research domains, publication years, and document formats. For each document in the dataset, we prepared:

- The original PDF file
- The extracted text (using a standardised extraction tool)
- A metadata file containing the nine predefined labels and their locations in the extracted text

The metadata annotations were created through a semi-automated process. For each document:

- Using the DOI from S2ORC, additional metadata was retrieved from CrossRef [33] via their API.
- PDFs were downloaded when available.
- Text was extracted from the first page of each PDF and normalized to eliminate irregularities such as inconsistent spacing and line breaks.
- Both exact and fuzzy matching techniques were employed to extract critical metadata elements (title, authors, abstract, etc.).
- For each identified metadata element, positions were determined based on their locations within the extracted text.

The extracted metadata underwent a simple human validation step to verify accuracy and correct any extraction errors, ensuring high-quality annotations

3.1.2. SSOAR Generated Multidisciplinary Vision Dataset (SSOAR-GMVD)

The SSOAR-GMVD[1] dataset contains approximately 44,000 papers with both German and English content. We began by randomly selecting 100 German scientific papers from publications available in the SSOAR repository[2], representing various layouts and styles. During the manual annotation phase, we identified the 28 most common layouts across these publications.

To expand our dataset, we developed an automated approach to generate synthetic papers based on these identified layouts. We randomly extracted metadata records from SSOAR, DBLP, and a list of scientific affiliations from Wikipedia. For each of the 28 common layouts, we generated an average of 1,600 synthetic papers by randomly inserting the extracted metadata at their corresponding positions on the first page of document templates. This approach allowed us to create a diverse and representative dataset while maintaining layout consistency with real-world scientific publications.

For the shared task evaluation, we utilized a subset of 8,518 documents from the complete collection, which were fully annotated with bounding box information. This subset contains over 2 million words, with approximately 1.6 million labeled words (79.2% of the total content). The average document in this subset contains 241 words, with 191 labeled for metadata extraction.

The class distribution in the annotated subset shows a predominance of abstract content (46.87% of all labeled words), followed by author names (3.75%), titles (4.97%), journal information (1.47%), and affiliations (0.96%). Structured elements like DOIs (0.08%) and email addresses (0.04%) constitute a smaller but critical portion of the dataset.

The SSOAR-GMVD dataset additionally includes affiliations in 8,242 documents (96.8%) and email addresses in 3,407 documents (40.0%). While affiliations were not included in the primary evaluation metrics, this additional element provides valuable information for institutional analysis and author disambiguation.

The dataset spans a wide temporal range, with publications dating from 1900 to 2020, though most documents (approximately 54%) were published between 2000-2010. The distribution across publishers shows significant diversity, with content from Nature (1.03%), SAGE (0.20%), and numerous specialized publishers in the social sciences. For evaluation purposes, the dataset was divided into training (70%), validation (15%), and testing (15%) splits.

[1] https://github.com/zeyd31/SSOR_GMVD
[2] https://www.gesis.org/en/ssoar/home

The SSOAR-GMVD dataset is particularly valuable for approaches that leverage computer vision techniques, as it provides pixel-level bounding box annotations for metadata elements. The dataset's focus on disciplines with challenging layout formats (Social Sciences, Humanities, Law, and Administration) makes it especially useful for testing the robustness of metadata extraction systems across diverse document templates.

4. Evaluation Methodology

The evaluation of submissions was designed to be comprehensive, incorporating multiple metrics to assess different aspects of metadata extraction performance:

- **Accuracy**: The proportion of metadata elements correctly identified as present or absent.
- **Precision**: The proportion of extracted metadata that correctly matched the gold standard.
- **Recall**: The proportion of gold standard metadata that was correctly extracted.
- **F1 score**: The harmonic mean of precision and recall, serving as the primary ranking metric.

A key feature of the evaluation was using Levenshtein Similarity to assess the match between extracted metadata and the gold standard. Rather than requiring exact matches, the evaluation considered an extraction correct if it achieved at least 95% Levenshtein Similarity with the reference text. This approach acknowledged the inherent challenges in extracting metadata from PDFs, where minor differences in whitespace, punctuation, or character encoding might occur without significantly affecting the semantic content. The evaluation function was provided to participants in the main repository to ensure transparency and to allow for consistent self-assessment during system development.

5. Task Timeline and Organization

The MESD shared task followed this timeline:

- Release of training datasets: January 27, 2025
- Release of testing datasets: February 15, 2025 (extended from original plan)
- Deadline for system submissions: March 4, 2025 (extended from February 25, 2025)
- Announcement of results: March 6, 2025 (extended from February 27, 2025)
- Paper submission deadline: March 6, 2025
- Notification of acceptance: April 3, 2025
- Camera-ready submission: April 17, 2025
- Workshop: June 1 or 2, 2025

The task was organized by a team from Fraunhofer FIT, Germany, with support from the Natural Language Processing community. The datasets, evaluation scripts, and submission instructions were made available through a dedicated repository, and participants were encouraged to contact the organizers with questions or clarifications. Participants were allowed to use either or both datasets for their system development, but the final evaluation was conducted on the test portions of both datasets to assess performance across different document types and annotation styles.

6. Future Directions

While detailed baseline performance on these datasets is not included in this paper, we refer readers to our companion work [4], which provides comprehensive evaluation of several approaches including established tools like GROBID, classical machine learning approaches, deep learning models and large language models. That analysis quantitatively demonstrates the challenges in metadata extraction

across different document types and layouts, particularly for multilingual content and documents from disciplines with less standardized formatting.

Based on our experience in designing the MESD shared task and analyzing the challenges in metadata extraction, we propose several considerations for future initiatives in this area:

- **Broader Task Definition**: Expanding the scope to include additional metadata elements or related tasks, such as citation extraction or bibliographic reference parsing, could increase the appeal and applicability of research in this domain.
- **Tiered Evaluation**: Implementing a tiered evaluation framework that acknowledges different levels of extraction difficulty could provide more nuanced assessment and better reflect the complexity of the task.
- **Hybrid Approaches**: Encouraging the development of methods that combine rule-based techniques with machine learning models might better address the variety of document formats and metadata representations.
- **Integration with Existing Workflows**: Aligning extraction techniques more closely with existing scholarly document processing workflows and tools could enhance their practical relevance and facilitate adoption of the resulting technologies.
- **Multimodal Approaches**: The SSOAR-GMVD dataset opens opportunities for exploring computer vision techniques in conjunction with text-based methods, potentially leading to more robust metadata extraction systems that can leverage both the visual and textual aspects of documents.
- **Cross-lingual Extensions**: Expanding the datasets to include non-English documents would address the important challenge of extracting metadata from multilingual scholarly literature.
- **FAIR Digital Objects**: Extracted metadata can serve as the foundation for creating FAIR Digital Objects (FDOs) of scholarly articles. FDOs represent a paradigm for making digital content machine-actionable through persistent identifiers, type definitions, and rich metadata. By transforming extracted metadata into standardized FDO representations, scholarly documents can be seamlessly integrated into knowledge graphs, research infrastructures, and scientific workflows. This would enhance not only the discoverability of articles but also enable automated reasoning and integration with computational tools, further advancing the FAIR principles in scholarly communication.

7. Conclusion

The MESD shared task was conceived to address an important challenge in scientific information management: the extraction of metadata from scholarly documents to enhance their findability and reusability. The task design, dataset creation, and evaluation methodology reflect the complexity and importance of this problem. The creation of two complementary datasets—S2ORC_Exp500v1 with detailed text annotations and SSOAR-GMVD with computer vision-oriented bounding box annotations—provides valuable resources for researchers approaching the problem from different angles. We believe these datasets will enable more robust comparisons between different methodological approaches to metadata extraction. The need for effective metadata extraction from scholarly documents remains pressing in the scientific community. We hope that the resources developed for the MESD shared task will contribute to ongoing research in this area. Future initiatives can build on these foundations to advance the state of the art in scholarly document processing and bring us closer to a fully FAIR scholarly ecosystem.

Declaration on Generative AI

While preparing this work, the authors used AI-assisting tools such as ChatGPT and Grammarly to check grammar and spelling, as well as paraphrase and reword. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] Z. Boukhers, N. Beili, T. Hartmann, P. Goswami, M. A. Zafar, Mexpub: Deep transfer learning for metadata extraction from german publications, in: 2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL), IEEE, 2021, pp. 250–253.

[2] Z. Boukhers, A. Bouabdallah, Vision and natural language for metadata extraction from scientific pdf documents: a multimodal approach, in: Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries, 2022, pp. 1–5.

[3] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al., The fair guiding principles for scientific data management and stewardship, Scientific data 3 (2016) 1–9.

[4] Z. Boukhers, C. Yang, Beyond feature learning: Textmap and its comparative performance against traditional methods and large language models for pdf metadata extraction, arXiv preprint arXiv:2501.05082 (2025).

[5] I. G. Councill, C. L. Giles, M.-Y. Kan, Parscit: an open-source crf reference string parsing package, LREC, Vol. 8. (2008) 661–667.

[6] Grobid (2008–2021). arXiv:1:dir:dab86b296e3c3216e2241968f0d63b68e8209d3c.

[7] M.-Y. Day, R. T.-H. Tsai, C.-L. Sung, C.-C. Hsieh, C.-W. Lee, S.-H. Wu, K.-P. Wu, C.-S. Ong, W.-L. Hsu, Reference metadata extraction using a hierarchical knowledge representation framework, Decision Support Systems 43 (2007) 152–167. URL: https://www.sciencedirect.com/science/article/pii/S0167923606001205. doi:https://doi.org/10.1016/j.dss.2006.08.006.

[8] A. Kawtrakul, C. Yingsaeree, A unified framework for automatic metadata extraction from electronic document, in: Proceedings of The International Advanced Digital Library Conference. Nagoya, Japan, 2005.

[9] D. Tkaczyk, P. Szostek, M. Fedoryszak, P. J. Dendek, L. Bolikowski, Cermine: Automatic extraction of structured metadata from scientific literature, Int. J. Doc. Anal. Recognit. 18 (2015) 317–335. URL: https://doi.org/10.1007/s10032-015-0249-8. doi:10.1007/s10032-015-0249-8.

[10] Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, S. Liu, Y. Zeng, S. Mehrabi, S. Sohn, et al., Clinical information extraction applications: a literature review, Journal of biomedical informatics 77 (2018) 34–49.

[11] F. Peng, A. McCallum, Information extraction from research papers using conditional random fields, Inf. Process. Manage. 42 (2006) 963–979. URL: https://doi.org/10.1016/j.ipm.2005.09.002. doi:10.1016/j.ipm.2005.09.002.

[12] A. Souza, V. Moreira, C. Heuser, Arctic: metadata extraction from scientific papers in pdf using two-layer crf (2014) 121–130.

[13] Z. Huang, W. Xu, K. Yu, Bidirectional lstm-crf models for sequence tagging, CoRR abs/1508.01991 (2015). URL: http://arxiv.org/abs/1508.01991.

[14] D. An, L. Gao, Z. Jiang, R. Liu, Z. Tang, Citation metadata extraction via deep neural network-based segment sequence labeling, in: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM '17, Association for Computing Machinery, New York, NY, USA, 2017, p. 1967–1970. URL: https://doi.org/10.1145/3132847.3133074. doi:10.1145/3132847.3133074.

[15] J. P. C. Chiu, E. Nichols, Named entity recognition with bidirectional lstm-cnns, CoRR abs/1511.08308 (2015). URL: http://arxiv.org/abs/1511.08308.

[16] P. R. Nayaka, R. Ranjan, An efficient framework for metadata extraction over scholarly documents using ensemble cnn and bilstm technique (2023) 1–9.

[17] C. G. Stahl, S. R. Young, D. Herrmannova, R. M. Patton, J. C. Wells, Deeppdf: A deep learning approach to extracting text from pdfs (2018). URL: https://www.osti.gov/biblio/1460210.

[18] X. Zhong, J. Tang, A. J. Yepes, Publaynet: largest dataset ever for document layout analysis (2019) 1015–1022.

[19] Z. Boukhers, N. Beili, T. Hartmann, P. Goswami, M. A. Zafar, Mexpub: Deep transfer learning for metadata extraction from german publications, in: 2021 ACM/IEEE Joint Conference on Digital

Libraries (JCDL), IEEE, 2021.
[20] D. Ali, K. Milleville, S. Verstockt, N. Van de Weghe, S. Chambers, J. M. Birkholz, Computer vision and machine learning approaches for metadata enrichment to improve searchability of historical newspaper collections, Emerald Publishing Limited, 2023.
[21] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980–2988. doi:10.1109/ICCV.2017.322.
[22] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2117–2125.
[23] R. Liu, L. Gao, D. An, Z. Jiang, Z. Tang, Automatic document metadata extraction based on deep networks, in: X. Huang, J. Jiang, D. Zhao, Y. Feng, Y. Hong (Eds.), Natural Language Processing and Chinese Computing, Springer International Publishing, Cham, 2018, pp. 305–317.
[24] V. Balasubramanian, S. G. Doraisamy, N. K. Kanakarajan, A multimodal approach for extracting content descriptive metadata from lecture videos, J. Intell. Inf. Syst. 46 (2016) 121–145. URL: https://doi.org/10.1007/s10844-015-0356-5. doi:10.1007/s10844-015-0356-5.
[25] M. K. Chandrasekaran, G. Feigenblat, D. Freitag, T. Ghosal, E. Hovy, P. Mayr, M. Shmueli-Scheuer, A. de Waard, Overview of the first workshop on scholarly document processing (sdp), in: Proceedings of the first workshop on scholarly document processing, 2020, pp. 1–6.
[26] Z. Boukhers, P. Mayr, S. Peroni, Bibliodap'21: The 1st workshop on bibliographic data analysis and processing, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021, pp. 4110–4111.
[27] S. Anzaroot, A. Mccallum, A new dataset for fine-grained citation field extraction, ICML Workshop on Peer Reviewing and Publishing Models. (2013).
[28] K. Seymore, A. Mccallum, R. Rosenfeld, Learning hidden markov model structure for information extraction, in: In AAAI 99 Workshop on Machine Learning for Information Extraction, 1999, pp. 37–42.
[29] H. Li, I. Councill, W.-C. Lee, C. L. Giles, Citeseerx: an architecture and web service design for an academic document search engine, in: Proceedings of the 15th international conference on World Wide Web, 2006, pp. 883–884.
[30] K. Lo, L. L. Wang, M. Neumann, R. Kinney, D. S. Weld, S2orc: The semantic scholar open research corpus, arXiv preprint arXiv:1911.02782 (2020).
[31] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.
[32] D. An, L. Gao, Z. Jiang, R. Liu, Z. Tang, Citation metadata extraction via deep neural network-based segment sequence labeling, in: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, 2017, pp. 1967–1970.
[33] G. Hendricks, D. Tkaczyk, J. Lin, P. Feeney, Crossref: The sustainable source of community-owned scholarly metadata, Quantitative Science Studies 1 (2020) 414–427.