Computer Science Foundations for Digital Libraries: Algorithms, Systems, and Applications

Donatella Firmani[1], Stefano Mizzaro[2], Beatrice Portelli[2,3], Gianmaria Silvello[4*], Sara Tonelli[5]

[1]Department of Statistical Sciences, Sapienza University of Rome, Piazzale Aldo Moro 5, Rome, Italy.
[2]Department of Mathematics, Computer Science and Physics, University of Udine, Via delle Scienze, 206, Udine, Italy.
[3]Department of Agriculture, Animal, Environmental and Food Sciences, University of Udine, Via delle Scienze, 206, Udine, Italy.
[4*]Department of Information Engineering, University of Padua, Via Gradenigo, 6/b, Padova, Italy.
[5]Department of Digital Humanities, Fondazione Bruno Kessler, Via Sommarive, 18, Trento, Italy.

*Corresponding author(s). E-mail(s): gianmaria.silvello@unipd.it;
Contributing authors: donatella.firmani@uniroma1.it; mizzaro@uniud.it; beatrice.portelli@uniud.it; satonelli@fbk.eu;

Abstract

Digital libraries face challenges in quality, accessibility, and usage of resources. This issue presents seven papers offering computational and technical solutions to these problems: data quality through validation and monitoring, AI evaluation of information systems, and enhanced content discoverability. Research also covers knowledge representation with new provenance models, deep learning for bibliographic control, metadata-driven access to underrepresented languages, and computational methods for restoring historical documents. These papers showcase how modern techniques like machine learning, semantic web technologies, knowledge graphs, and image processing tackle digital library challenges, improving resource quality and accessibility.
These papers were selected from the 21st Italian Research Conference on Digital Libraries (IRCDL 2025), held in Udine, Italy, on 20–21 February 2025, which has served since 2005 as a key annual forum bringing together researchers from academia, government, and industry to address topics spanning computer science, digital humanities, information science, librarianship, archival science, museum studies, and cultural heritage.

Keywords: Scholarly Document Processing, Text Analytics, Citation Behaviours, Author Name Disambiguation, Recommendation System, Open Access, Web Archiving

Preface

Validating and monitoring bibliographic and citation data in OpenCitations collections by Heibi, Peroni, and Rizzetto [7] addresses the challenge of maintaining high-quality open research information within the OpenCitations infrastructure by introducing tools for validating and monitoring bibliographic metadata and citation data. The paper develops a custom validation tool tailored to the OpenCitations Data Model, designed to detect and explain ingestion errors from heterogeneous sources, whether due to upstream data inconsistencies or internal software bugs, alongside a quality monitoring tool to track known data issues post-publication. These tools were applied in two scenarios: validating metadata and citations from Matilda,[1] a potential future source, and monitoring data quality in the existing OpenCitations Meta dataset. The validation tool successfully identified a variety of structural and semantic issues in the Matilda dataset, demonstrating its precision, while the monitoring tool enabled the detection and quantification of recurring problems in the OpenCitations Meta collection. Together, these tools proved effective in enhancing the reliability of OpenCitations' published data, representing a step toward ensuring high-quality bibliographic data in open research infrastructures, though currently limited to the data model adopted by OpenCitations. This paper is the extended version of the original one published in IRCDL 2025 [9].

Toward purpose-oriented topic model evaluation enabled by large language models by Tan and D'Souza [11] presents a framework for automated evaluation of dynamically evolving topic models using Large Language Models, addressing the limitations of widely used automated metrics such as coherence and diversity that often capture only narrow statistical patterns and fail to explain semantic failures in practice. The study introduces a purpose-oriented evaluation framework that employs nine LLM-based metrics spanning four key dimensions of topic quality: lexical validity, intra-topic semantic soundness, inter-topic structural soundness, and document-topic alignment soundness. The framework is validated through adversarial and sampling-based protocols and applied across datasets spanning news articles, scholarly publications, and social media posts, using multiple topic modeling methods and open-source LLMs. The analysis demonstrates that LLM-based metrics provide interpretable, robust, and task-relevant assessments, uncovering critical weaknesses in topic models such as redundancy and semantic drift that are often missed by traditional metrics, thereby supporting the development of scalable, fine-grained evaluation tools for maintaining topic relevance in dynamic datasets relevant to digital library systems. This paper is the extended version of the original one published in IRCDL 2025 [10].

Searching Agricultural Learning Experiences in the Metaverse via Textual and Visual Queries Within the AgriMus project by Abdari, Falcon, and Serra [2] introduces the AgriMus project, which aims to design agricultural-themed museums in the Metaverse covering a broad range of topics to support agricultural education through immersive virtual environments. The paper addresses the challenge of finding suitable educational spaces for specific themes by creating a dataset of 83 AgriMuseums, each covering a specific agricultural topic, where users can explore and retrieve specific educational experiences through visual or textual queries. The proposed hierarchical method models these virtual museums by progressively integrating information extracted from visual data into rooms and finally into representations for the whole exhibition, with validation conducted in a zero-shot setting on the collected data. The approach achieves up to 27.23% R@1 and 41.33 MRR for textual queries, and 47.91% R@1 and 59.54 MRR for visual queries, demonstrating the effectiveness of vision-language models in supporting agricultural education through metaverse experiences and enhancing content discoverability in specialized virtual learning environments. This paper is the extended version of the original one published in IRCDL 2025 [1].

Provenance-driven nanopublications: representing source lineage and trust networks for multi-source assertions by Menotti, Marchesin, Giachelle, and Silvello [8] extends the nanopublication model with knowledge provenance to support multi-source assertions, addressing the limitation that standard nanopublications are designed to represent assertions from

[1] https://doi.org/10.5281/ZENODO.15224594

a single evidence source and lack the capacity to express that an assertion is derived from multiple, potentially conflicting sources. The paper introduces an extended nanopublication model with a fourth named graph dedicated to capturing knowledge provenance, representing different sources that support or conflict with a given assertion, and develops the PROV-K ontology to represent provenance information for multi-source assertions, built upon PROV-O and grounded in knowledge provenance literature. The authors serialize 197,511 extended nanopublications representing gene expression-cancer associations generated by the Collaborative Oriented Relation Extraction (CORE) System, integrating them into the CoreKB platform.[2] Additionally, a trust network comprising 45,649 facts and multiple LLM agents is constructed to demonstrate trust relationships capabilities, providing insights into fact reliability and uncertainty regions within knowledge graphs, thereby advancing the representation and management of provenance for aggregated scientific claims in digital libraries. This paper is the extended version of the original one published in IRCDL 2025 [8].

Digital Maktaba project: Toward a metadata-driven, LLM-assisted framework for arabic digital libraries by El Ganadi, Gagliardelli, and Ruozzi [5] proposes a metadata-driven framework for Arabic-script digital libraries, addressing distinct challenges related to access, discoverability, and long-term preservation presented by the rapid digitization of cultural heritage collections featuring Arabic-script texts. The framework leverages validated metadata from the Diamond catalogue and the La Pira Library's extensive collection, employing frontispiece images and the Kraken OCR engine within the eScriptorium platform to train high-accuracy recognition models that address the technical complexities of Arabic script, including calligraphy, diacritics, and ligatures. The cataloging workflow is structured around international standards such as Dublin Core and informed by both topographic and thematic classification practices, while evaluating the use of large language models, including Arabic-specialized models, to assess their potential for extracting semantic metadata from digitized texts. By combining human-validated metadata with machine learning pipelines, the Digital Maktaba project aims to provide a scalable, standards-aligned approach for building Arabic digital libraries with broader applicability to other underrepresented language collections, though current LLM-based topic extraction results show that zero-shot prompting consistently outperforms few-shot configurations, with Qwen2.5 achieving 63.93% accuracy and GPT-4.1-mini reaching 78.33%. This paper is the extended version of the original one published in IRCDL 2025 [6].

Deep Learning Approaches to Author Name Disambiguation: A Survey by Cappelli, Colavizza, and Peroni [3] provides a systematic review of state-of-the-art author name disambiguation techniques based on deep learning within the timeframe 2016-2024, addressing the critical task for digital libraries of linking existing authors with their respective publications in the face of challenges from homonymy, synonymy, and lack of persistent identifiers. The methodology includes a systematic literature review using Google Scholar with keywords related to author name disambiguation and deep learning, yielding 52 documents of which 28 were selected after full-text assessment, with approaches categorized based on learning paradigms: supervised, unsupervised, and hybrid methods. The analysis reveals that supervised methods predominantly focus on author assignment tasks, while unsupervised approaches excel at author grouping, and hybrid approaches demonstrate superior performance, with the top-performing method reaching 89.7% F1-score on AMiner datasets. However, heavy reliance on AMiner datasets and lack of standardized evaluation frameworks remain critical challenges for the field's advancement, with the paper also including bibliometric analysis using OpenCitations data with network visualization via Gephi, identifying influential works and research communities in the domain. This paper is the extended version of the original one published in IRCDL 2025 [4].

AI-driven enhancement of historical documents by Ziran, Mecella, and Marinai [12] investigates image processing algorithms and deep learning models for enhancing historical documents from Late Antiquity to the early Middle Ages that suffer from degraded image quality due

[2] https://gda.dei.unipd.it/

to aging, inadequate preservation, and environmental factors, presenting significant challenges for paleographical analysis. The research focuses on documents containing crucial graphical symbols representing administrative, economic, and cultural information, which are time-consuming and error-prone to interpret manually, by implementing three distinct enhancement pipelines: Model A using edge detection and superpixel segmentation, Model B emphasizing morphological noise suppression, and a deep learning approach employing Faster R-CNN with GAN-based synthetic data generation for symbol-aware detection and reconstruction. Quantitative evaluation on 100 historical document images demonstrates that Model A achieves superior performance in edge preservation and contrast enhancement with a PSNR of 32.4 dB and SSIM of 0.89, Model B excels in noise reduction while maintaining edge preservation with EPI of 0.85, and the deep learning approach shows 73% accuracy in reconstructing damaged symbols, particularly excelling in areas where traditional methods struggle, consistently outperforming the existing Hierax approach and contributing to improving the quality of historical document analysis for graphical symbol interpretation in paleographical studies. This paper is the extended version of the original one published in IRCDL 2025 [13].

Overall, these seven papers represent a revised and extended version of the original papers published in the Proceedings of the 21st Italian Research Conference on Digital Libraries (IRCDL 2025) held in Udine, Italy, on 20–21 February 2025.[3] The papers were selected among the best full papers published in IRCDL 2025. Since 2005, IRCDL[4] has served as a key annual forum on digital libraries and associated issues, bringing together researchers from academia, government, and industry to address topics spanning computer science, digital humanities, information science, librarianship, archival science, museum studies, and cultural heritage, with the 2025 edition featuring two distinct tracks on Computer Science Foundations for Digital Libraries and Digital Humanities.

[3] https://ircdl2025.uniud.it/
[4] http://ims.dei.unipd.it/websites/ircdl/home.html

References

[1] Abdari, A., Falcon, A., and Serra, G. (2025a). Agrimus: Developing museums in the metaverse for agricultural education. In Cornia, M., Di Nunzio, G. M., Firmani, D., Mizzaro, S., Serra, G., Tonelli, S., and Tremamunno, A., editors, *Proceedings of the 21st Conference on Information and Research science Connecting to Digital and Library science, Udine, Italy, February 20-21, 2025*, volume 3937 of *CEUR Workshop Proceedings*. CEUR-WS.org.

[2] Abdari, A., Falcon, A., and Serra, G. (2025b). Searching agricultural learning experiences in the metaverse via textual and visual queries within the agrimus project. *International Journal on Digital Libraries*, 26(4):20.

[3] Cappelli, F., Colavizza, G., and Peroni, S. (2025a). Deep learning approaches to author name disambiguation: A survey. *International Journal on Digital Libraries*, 26(4):21.

[4] Cappelli, F., Colavizza, G., and Peroni, S. (2025b). Recent developments in deep learning-based author name disambiguation. In Cornia, M., Di Nunzio, G. M., Firmani, D., Mizzaro, S., Serra, G., Tonelli, S., and Tremamunno, A., editors, *Proceedings of the 21st Conference on Information and Research science Connecting to Digital and Library science, Udine, Italy, February 20-21, 2025*, volume 3937 of *CEUR Workshop Proceedings*. CEUR-WS.org.

[5] El Ganadi, A., Gagliardelli, L., and Ruozzi, F. (2025). Digital maktaba project: Toward a metadata-driven, llm-assisted framework for arabic digital libraries. *International Journal on Digital Libraries*, 26(4):19.

[6] Ganadi, A. E., Gagliardelli, L., Aftar, S., and Ruozzi, F. (2025). Digital maktaba project: Proposing a metadata-driven framework for arabic library digitization. In Cornia, M., Di Nunzio, G. M., Firmani, D., Mizzaro, S., Serra, G., Tonelli, S., and Tremamunno, A., editors, *Proceedings of the 21st Conference on Information and Research science Connecting to Digital and Library science, Udine, Italy, February 20-21, 2025*, volume 3937 of *CEUR Workshop Proceedings*. CEUR-WS.org.

[7] Heibi, I., Peroni, S., and Rizzetto, E. (2025). Validating and monitoring bibliographic and citation data in opencitations collections. *International Journal on Digital Libraries*, 26(3):16.

[8] Menotti, L., Marchesin, S., Giachelle, F., and Silvello, G. (2025). Provenance-driven nanopublications: representing source lineage and trust networks for multi-source assertions. *International Journal on Digital Libraries*, 26(4):24.

[9] Peroni, S. and Rizzetto, E. (2025). A tool for validating and monitoring bibliographic data in open research information systems: The opencitations collections. In Cornia, M., Di Nunzio, G. M., Firmani, D., Mizzaro, S., Serra, G., Tonelli, S., and Tremamunno, A., editors, *Proceedings of the 21st Conference on Information and Research science Connecting to Digital and Library science, Udine, Italy, February 20-21, 2025*, volume 3937 of *CEUR Workshop Proceedings*. CEUR-WS.org.

[10] Tan, Z. and D'Souza, J. (2025a). Bridging the evaluation gap: Leveraging large language models for topic model evaluation. In Cornia, M., Di Nunzio, G. M., Firmani, D., Mizzaro, S., Serra, G., Tonelli, S., and Tremamunno, A., editors, *Proceedings of the 21st Conference on Information and Research science Connecting to Digital and Library science, Udine, Italy, February 20-21, 2025*, volume 3937 of *CEUR Workshop Proceedings*. CEUR-WS.org.

[11] Tan, Z. and D'Souza, J. (2025b). Toward purpose-oriented topic model evaluation enabled by large language models. *International Journal on Digital Libraries*, 26(4):23.

[12] Ziran, Z., Mecella, M., and Marinai, S. (2025a). Ai-driven enhancement of historical documents. *International Journal on Digital Libraries*, 26(4):22.

[13] Ziran, Z., Mecella, M., and Marinai, S. (2025b). Enhancing historical documents: Deep learning and image processing approaches. In Cornia, M., Di Nunzio, G. M., Firmani, D., Mizzaro, S., Serra, G., Tonelli, S., and Tremamunno, A., editors, *Proceedings of the 21st Conference on Information and Research science Connecting to Digital and Library science, Udine, Italy, February 20-21, 2025*, volume 3937 of *CEUR Workshop Proceedings*. CEUR-WS.org.