MAGIC: A Computational Framework for Large-Scale Manuscript Digitization, Preservation, and Semantic Enrichment

Adriano Ettari
Depart. of Physics
University Federico II
Naples, Italy
adriano.ettari@unina.it

Guido Russo
Depart. of Physics
University Federico II
Naples, Italy
guido.russo@unina.it

Abstract

This work presents a computationally oriented framework for large-scale manuscript digitization, preservation, and analysis, addressing thematic and methodological challenges in the digital humanities. The work is based on the 3-year experience with the MAGIC project [Russo et al., 2020], [Conte et al., 2024] at the University of Naples Federico II, in Italy, a project which has involved 12 young researchers under the supervision of two expert professors. The project is now going to finish at end 2026, after having digitized and analyzed over 250 books and over 2000 parchments, as well as archival documents. Most of the books date to the medieval age. All the work done is published on the web site www.magic.unina.it.

The proposed workflow integrates acquisition, storage, metadata formalization, format standardization, and post-processing into a cohesive architecture designed to support interdisciplinary humanities research while ensuring long-term data sustainability and accessibility, and preservation.

High-resolution digitization [International Federation of Library Associations and Institutions (IFLA), 2015], [Andreoli et al., 2022] is achieved through specialized planetary scanning systems, deployed within an infrastructure supported by the University data center, which provides petabyte-scale storage and high-speed networking. The scale of acquired data needs formalized data management strategies that allow efficient storage, retrieval, and computational processing of multimodal manuscript representations.

Knowledge representation is implemented through the integration of descriptive (to describe and support retrieval) and structural (to emphasize relationships among volumes) metadata, modeled according to METS and MAG standards [Feliciati, 2009], version 2.01. XML-based metadata schemas encode the content, the structure, the cultural background, and the conservation status of the books, thereby supporting semantic interoperability and advanced retrieval mechanisms. The adoption of the FITS [Wells et al., 1981] format introduces a self-documenting data structure combining extensible metadata headers with numerical image arrays. The Header contains metadata such as image dimensions and the pixel data type and is editable by the user; the image array contains the image itself and can be bidimensional or bigger (i.e. stacked images). The conversion from TIFF to FITS facilitates long-term preservation, reproducibility, and extensible annotation capabilities within digital humanities corpora. The addition of FITS is not new, it was already introduced by [**?**] and, as Allegranza wrote, FITS guarantees more than 30 years of usage and long term compatibility.

Computational post-processing includes machine learning-based segmentation models [Ettari et al., 2025] to isolate textual and ornamental features from the background of the page which may be affected by degradation effects. An example of such effect we dealt with is bleed-through. It consists of mirror-image

brown stains on the reverse side of the page. Old ink penetrates into parchment fibers, and over time the oxidation of the ferrous sulfate contained in the ink generates acids that cause the bleed-through effect. Once we isolated the good parts of the pages (i.e. ornaments and text), we reconstructed the background with inpainting techniques, improving both visual legibility and analytical accessibility. These methods were experimentally applied to ancient manuscripts, like several copies of Divine Comedy by Dante Alighieri, demonstrating how computational vision techniques can enhance paleographic and codicological analysis. Evaluating quantitatively the results is not straightforward, since ground-truth cleaned images are not available. We are working to manually recreate a substantial number of bleed-through-free manuscript images in order to evaluate our methods, and we plan to publish this evaluation work as a separate article soon.

Moreover, the project explores advanced optical character recognition pipelines tailored to early printed materials and incunabula [Momtaz et al., 2025]. Beyond improving transcription quality, this work contributes significantly to the diffusion of Digital Humanities by facilitating the large-scale digitization, accessibility, and computational analysis of cultural heritage collections. Conventional OCR engines often produce transcriptions of historical texts with character error rates (CER) exceeding 20–25%, severely limiting the utility of digitized sources for scholarly research, full-text search, metadata extraction, and public access. To address these challenges, we employed Kraken, a neural network-based OCR toolkit, trained on manually annotated PAGE XML files containing region boundaries, baselines, and text line coordinates. This custom-trained segmentation model accurately extracts individual text lines from complex historical layouts, providing clean line-level inputs for subsequent recognition. Following segmentation, we applied TrOCR (medieval-data/trocr-medieval-print), a transformer-based OCR model pretrained on early printed texts. This model generates initial transcriptions while preserving historical spelling variations and character-level structures characteristic of incunabula. The OCR outputs were subsequently refined through a transformer-based post-correction stage, enabling substantial improvements in transcription quality. The complete pipeline reduced the Character Error Rate from approximately 38% in the raw OCR output to about 15%, corresponding to a relative error reduction of more than 50%, while nearly doubling the Bilingual Evaluation Understudy (BLEU) score from about 22 to 44 [Momtaz et al., 2025]. These improvements produce more reliable digital texts that can be effectively indexed, searched, and analyzed using computational methods, thereby supporting a wide range of Digital Humanities applications, including corpus linguistics, historical language studies, text mining, and semantic analysis. Furthermore, the modular and open-source architecture of the pipeline promotes interoperability, reproducibility, and long-term sustainability, lowering the barriers for libraries, archives, and cultural institutions seeking to digitize and disseminate historical collections. These techniques were evaluated using materials such as Memorie della Regale Accademia ercolanese di archeologia from the Pontaniana Academy, as well as sheet music for Tristan und Isolde by Richard Wagner, highlighting the adaptability of computational models across textual genres, documentary typologies, and historical periods.

- Employing the FITS format for long-term digital preservation, thereby reducing the need for future re-digitization efforts. By integrating descriptive, technical, and preservation metadata into a single archival object, the resulting digital representation becomes not merely a reproduction of the original manuscript, but a scientifically documented dataset that preserves both

the cultural artifact and its contextual information.

- Improving the accessibility and usability of historical manuscripts through automated bleed-through removal. The enhanced images provide a clearer and more readable representation of the original sources, facilitating consultation by students, non-specialist audiences, and cultural heritage practitioners, while also supporting expert scholarly analysis.
- Providing machine-readable transcriptions of manuscripts and early printed materials that enable full-text search, information retrieval, and corpus-based research. These transcriptions support a broad range of Digital Humanities applications, including text mining, linguistic analysis, semantic annotation, digital editions, and the creation of audiobooks and educational resources.

The paper will first analyze and discuss the state-of-the-art in the field, and then present, in a scientific way, the work done by our research group and the results we obtained, which represent a step forward in the diffusion of digital humanities.

References

Guido Russo, Luciano Aiosa, Giancarlo Alfano, Angelo Chianese, Fabio Cornevilli, Gian Marco Di Domenico, Pasqualino Maddalena, Andrea Mazzucchi, Ciro Muraglia, Felice Russillo, et al. Magic: Manuscripts of girolamini in cloud. In *IOP Conference Series: Materials Science and Engineering*, volume 949, page 012081. IOP Publishing, 2020.

Stefania Conte, Gian Marco Di Domenico, Andrea Mazzei, Andrea Mazzucchi, Guido Russo, Alessandro Salvi, Augusto Tortora, et al. The magic project: first research results. In *IRCDL*, pages 87–93, 2024.

International Federation of Library Associations and Institutions (IFLA). Guidelines for planning the digitization of rare book and manuscript collections, 2015. URL https://www.ifla.org/files/assets/rare-books-and-manuscripts/rbms-guidelines/guidelines-for-planning-digitization.pdf.

Lorisa Andreoli, Marina Cimino, and Gianluca Drago. Linee guida sulla digitalizzazione di documenti bidimensionali. *Università degli Studi di Padova—Sistema Bibliotecario di Ateneo*, 2022.

Pierluigi Feliciati. Reference schema mag 2.01 (metadati amministrativi e gestionali). Technical report, ICCU - Istituto Centrale per il Catalogo Unico delle Biblioteche Italiane, 2009. URL http://eprints.rclis.org/13400/.

Donald C Wells, Eric W Greisen, and Ronald H Harten. Fits: A flexible image transport system. In *Applications of Digital Image Processing to Astronomy*, volume 264, pages 298–315. SPIE, 1981.

Adriano Ettari, Massimo Brescia, Stefania Conte, Yahya Momtaz, and Guido Russo. Minimizing bleed-through effect in medieval manuscripts with machine learning and robust statistics. *Journal of Imaging*, 11(5):136, 2025.

Yahya Momtaz, Lorenza Laccetti, and Guido Russo. Modular pipeline for text recognition in early printed books using kraken and byt5. *Electronics*, 14(15): 3083, 2025.

Acknowledgments

This research was part of the MAGIC project, funded by the Ministry of Enterprise 426 and Made in Italy, for the creation of a prototype of a service center for technologies applied to cultural heritage. The project was funded by the Ministry of Enterprise and Made in Italy, (code n. F/130093/03/X38 and CUP: B69J23000560005).