HSPI The Open Access Publisher

JAIRI ISSN 3069-762X JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH AND INNOVATION

Short communication

# Artificial Intelligence in the Digital Humanities

**Guido Russo*, Yahya Momtaz and Sohail Maqsood**

Department of Physics, University of Naples "Federico II", Italy

## Abstract

The research in the field of Digital Humanities (DH) has so far been limited to digitization techniques and database organization. However, over the last years, many researchers have pointed out the need for an interdisciplinary approach, which goes from chemical and physical analyses, to microbiology, to Artificial Intelligence. In this paper, we summarize the efforts in this direction of our "MAGIC" project, active for 3 years now in the field of DH, in particular for the application of AI to DH to manuscripts and books of the period 1300-1600. The present short paper will deal with two problems: i) digital restoration of deteriorated pages of books and manuscripts, and ii) transcription of books and manuscripts via OCR and HTR techniques, and for both, we applied AI techniques to improve the automated process. It has to be remembered that only an interdisciplinary approach can give a full overview of these books and manuscripts; on the website www.magic.unina.it, we present the full project and the other work that is underway.

**More Information**

***Address for correspondence:**
Guido Russo, Department of Physics, University of Naples "Federico II", Italy,
Email: prof.guido.russo@gmail.com

https://orcid.org/0000-0001-5823-4393

**Submitted:** August 07, 2025
**Approved:** August 16, 2025
**Published:** August 18, 2025

**How to cite this article:** Russo G, Momtaz Y, Maqsood S. Artificial Intelligence in the Digital Humanities. J Artif Intell Res Innov. 2025; 1(1): 042-045. Available from: https://dx.doi.org/10.29328/journal.jairi.1001006



**Keywords:** Digital humanities; Ancient books; Image restoration; Manuscript transcription

Check for updates

OPEN ACCESS

## Introduction

Artificial Intelligence (AI) has also entered the world of Digital Humanities [1], particularly with machine learning and natural language processing. AI has forced researchers to change their approach to data analysis and data organization. Many projects around the world deal with DH nowadays, and many of them have started applying AI techniques [2].

The MAGIC project [3,4], being carried out in Napoli, Italy, is devoted to manuscripts and early printed books (from the year 1300 up to the year 1600), and since the beginning, it has been based on a multidisciplinary approach. Most of the material is suffering from aging problems, and many books have not been properly maintained over the centuries. It is important to understand which are the causes of such degradation to take the proper countermeasures: we are working with an FT-IR spectrophotometer to analyze paper contaminants, as an example. But it is also important to understand how to restore the original form and appearance of the books, at least in the digital world. Our group is coping with several aspects of the Digital Humanities, and in two areas we think to have introduced a significant contribution to the introduction of AI in the field, namely the minimization of the bleed-through effect and in the OCR and HTR for transcribing ancient documents. This short paper will summarize both. It has to be remembered that only an interdisciplinary approach can give a full overview of these books and manuscripts; we therefore have also carried out other parallel research, like the analysis of paper fragments with infrared spectrophotometry, and the DNA analysis of the mold found on the pages. These themes will be analyzed in a subsequent paper.

## Medieval manuscripts and books—minimizing the bleed-through effect with AI

Manuscripts and books dated from 1200 to 1600 are the main goal of our research. With ages ranging from 8 to 5 centuries, they all suffer from an unavoidable physical deterioration, and most of them are not accessible to readers, because of the fear of further deterioration. Many projects are underway to digitize these volumes, but our work has, in addition to digitization, tried the digital restoration of the book images (scanned pages at 400 dpi), to improve readability for all readers, especially in high schools and universities. One approach we worked on is the minimization of the problems due to the ink used [5-9].

Medieval manuscripts suffer from the so-called bleed-through effect: ink on the paper, very acidic at that time, has passed through the paper and is visible on the other side of the sheet, thus often making it difficult to read the content. This is particularly true for images (miniatures) in two or three colors, which were put at the beginning of chapters ("capolettera" or "drop cap"). What our research group has done during the last

couple of years is to develop a code, based on AI techniques, to minimize the bleed-through effect. In a recent paper [10], we have described the mathematics the code is based and the neural network algorithm. The importance of bleed-through removal lies in the possibility for students and young researchers to understand the text directly from the images, to compare different sources, and to look at the evolution of the books in their details. Depending on the difference between bleed-through-affected pixels and text-only pixels, one can choose between two methods: a "light removal" method and a "strong removal" method. Figure 1 shows the original (a, left) and corrected (b, right) image.

The code runs on NVIDIA GPU boards (we used both L40S and RTX5090 on PCs with i9 processors and 128 GB RAM) and takes from 5 to 10 seconds per page, as compared to 70 seconds without a GPU. The advantage of the method is that it is a "blind" method; that is, it can be run in an unattended way, without user interaction, and a page at a time: other methods require simultaneous analysis of the *recto* and *verso* of the sheet, with *de visu* alignment. If you want to minimize the bleed-through effect on a single manuscript, you have from 400 to 500 pages, and manual activities have to be avoided. In our case, we have about 200 manuscripts of Dante Alighieri's *Divine Comedy*.

We have now just started a study to determine a metric to be adopted, able to compare the results of various methods in a scientific way, thus allowing the user to adopt their criteria for bleed-through minimization. The work is not complete yet, but basically, we use a two-parameter model, thus creating a virtual Magic Quadrant. According to the user's needs, one can choose to remain in the lower right corner and accept that images in the book are not well restored to their original form, and this may be the case for a philology researcher. But another user may choose to study in detail the images rather than the text, which could be well known (we have access to digitized images of about 200 different copies of Dante Alighieri's Divine Comedy, whose text is now well known), and in this case, he will remain in the upper left corner of the quadrant. The work is underway and will be the base for an article shortly (Figure 2).

An example of pages with a strong bleed-through and figures ("illuminations") is given below (a: Madrid, Biblioteca Nacional de España, VITR/23/3 - Copyright of the image: Department of Humanities, University Federico II; b) Commentarij a Philippo Beroaldo conditi in asinum aureum Lucij Apuleij, Accademia Pontaniana, Napoli, - Copyright of the image: Department of Physics, University Federico II) (Figure 3).

## Optical Character Recognition (OCR) and Handwritten Text Recognition (HTR) in ancient books with AI

In several studies, it is necessary to make a transcription of an entire ancient book into textual files, e.g., in Word.

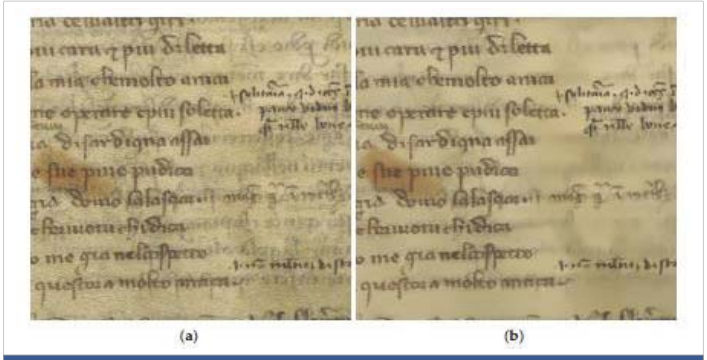


**Figure 1:** Bleed-through minimization: The original (a, left) and corrected (b, right) image.

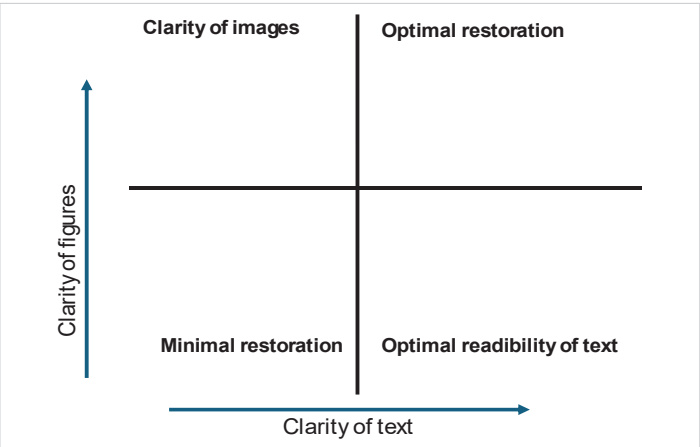


**Figure 2:** The Magic Quadrant of Image Optimal Restoration.

With printed books, the job is difficult but not impossible: early printed books like "incunabula" (years 1453-1500) use antique typeface characters, which may be irregular, and have complex layouts; in addition, the paper they are printed on suffers from the same bleed-through effect as coeval manuscripts. There exist computer applications, like the AI-based Transkribus tool [11], which can be of valuable help in this field. However, the result is not perfect, as the high Character Error Rate (CER) (misunderstood words) leads to a careful visual inspection of each transcribed page for manual corrections. In our approach to ancient books, we had the request to make the text of some books "searchable". Of course, just using Acrobat-like tools on the images is not enough- the results are much, much worse than what can be obtained on modern book images, for the above-mentioned reasons. What you need is therefore a "transcription" of the book. We devoted our efforts to two different problems: transcribing a printed book and transcribing a handwritten book, and in both cases, AI is of valuable help.

In a recent paper [12], we presented a modular pipeline that addresses these problems by combining modern layout analysis and language modeling techniques. Experimental results on a couple of 15th-century incunabula show a reduction in the CER from around 38% to around 15% with respect to other methods and an increase in the BLEU metric score from

a) b)

**Figure 3:** Digitized images of a page of Dante Alighieri's Divine Comedy (end 14th century), and of an *Incunabulum* (end 15th century), both used in our study.

22 to 44, confirming the effectiveness of our approach. This work demonstrates the potential of integrating transformer-based correction with layout-aware segmentation to enhance OCR accuracy in digital humanities applications.

To be more technical, the pipeline begins with the "Kraken" tool [13], a neural network-based tool tailored for early typographic structures. The first step of OCR is performed with Kraken's recognition engine, but we apply a post-correction using a fine-tuned "ByT5" transformer model [14] trained on manually aligned line-level data. By learning to map noisy OCR outputs to verified transcriptions, the model substantially improves recognition quality. The pipeline may start with a preprocessing step based on our previous work on bleed-through removal using robust statistical filters, including non-local means, Gaussian mixtures, biweight estimation, and Gaussian blur. This step enhances the legibility of degraded pages before OCR. The usage of AI lies in the learning phase based on a verified transcription of part of the text. The diagram below shows the pipeline steps in detail [12] (Figure 4).

But HTR on manuscripts is certainly more difficult than OCR. Transkribus allows one to develop a model and to train the model with visually recognized text: in general, one needs to transcribe the first 10% of the entire manuscript and then train the model on these pages iteratively. Only afterwards, Transkribus may be applied automatically to the entire manuscript, but it always requires a visual inspection and correction. The main problem is that in each manuscript, there are often at least 3-4 persons who wrote the book, and the training process requires restarting once the operator

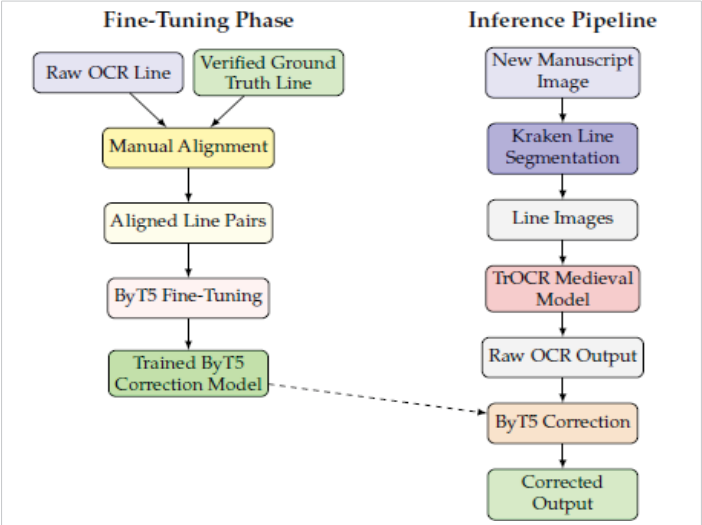


**Figure 4:** The pipeline steps in detail for OCR [12].

recognizes that the scribe has changed. We worked, therefore, on a manuscript (named Platea 3558) coming from the State Archive in Caserta, Italy, a 550-page manuscript very large in size (about 40x70 cm) dated 1826 that has a lot of tables and is full of historical information about the Borboni's kingdom in Naples [in the Figure 5: a) cover page; b) dedication; c) index].

Why do we need a transcription? Because we need a searchable document, not an image, and many of these documents in the State Archive in Caserta, within the famous historical monument "Reggia di Caserta," contain legal information about properties, devices, and similar things, which need to be accessed as they are still valid acts.

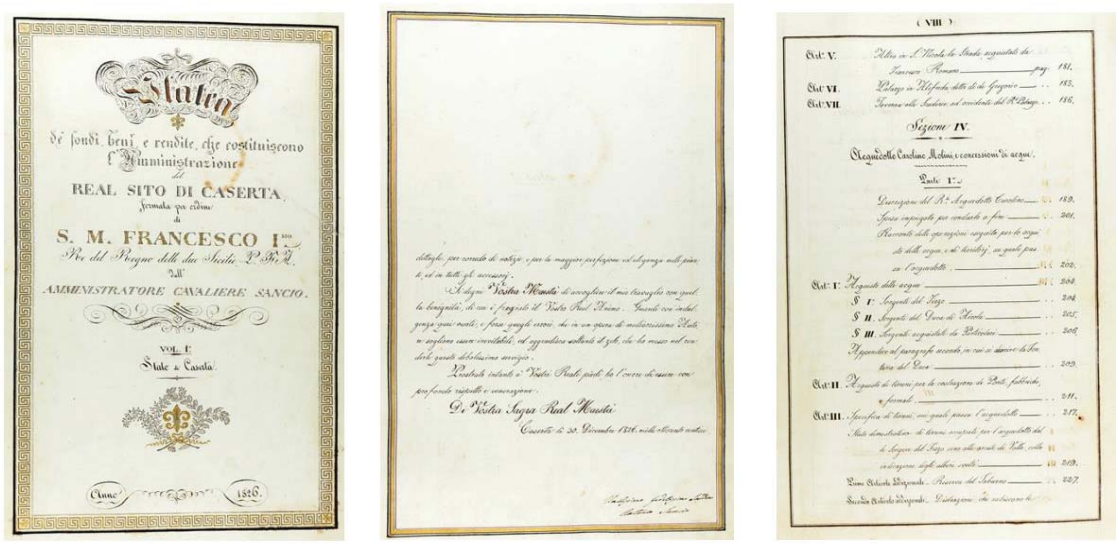


**Figure 5:** Full of historical information about the Borboni's kingdom in Naples a) cover page; b) dedication; c) index].

We are continuing the MAGIC project with several other activities, hoping to make a contribution to modern digital humanities studies.

## Acknowledgement

We thank the Director General of State Archives, Dr. Antonio Tarasco, for granting access to the State Archive in Caserta, and the Director of the State Archive in Caserta, Dr. Fortunata Manzi, for her valuable help in supporting our staff in transcribing the book, and for useful discussions about the application of AI to Digital Humanities. We also thank Dr. Stefania Conte, who organized most of the project activities for the first two years.

## References

1. Allegrezza S. Artificial intelligence as a key to the valorisation and conservation of cultural heritage: an interdisciplinary approach. Workshop AI, Cultural Heritage and Art, CINECA; Bologna, September 2023. Available from: https://dx.doi.org/10.1388/IIWORKSHOPAIBC
2. Caruso M, Spadaro A. Digital Humanities and Artificial Intelligence: An Accelerationist Perspective of the Future. Una Quantum 2022: Open Source Technologies for Cultural Heritage, Cultural Activities and Tourism. Proceedings. 2024;96(10). Available from: https://doi.org/10.3390/proceedings2024096010
3. Conte S, Maddalena PM, Mazzucchi A, Merola L, Russo G, Trombetti G. The Role of Project MA.G.I.C. In the Context of the European Strategies for the Digitization of the Library and Archival Heritage. In: Bucciero A, Fanini B, Graf H, Pescarin S, Rizvic S, editors. Proceedings of the Eurographics Workshop on Graphics and Cultural Heritage. Lecce, Italy: The Eurographics Association; 2023. Available from: https://doi.org/10.2312/gch.20231167
4. Conte S. Magic, a Service Center for Technologies Applied to Manuscripts and Printed Books. Uman Digit. 2025;9(20):661–684. Available from: https://doi.org/10.6092/issn.2532-8816/21038
5. Hanif M, Tonazzini A, Savino P, Salerno E, Tsagkatakis G. Document Bleed-Through Removal Using Sparse Image Inpainting. In: Proceedings of the 13th IAPR International Workshop on Document Analysis Systems (DAS); 2018 Apr 24–27; Vienna, Austria. Piscataway (NJ): IEEE; 2018;281–6. Available from: http://dx.doi.org/10.1109/DAS.2018.21
6. Dubois E, Pathak A. Reduction of bleed-through in scanned manuscript documents. In: Proceedings of the IS&T Conference on Image Processing, Image Quality, Image Capture Systems; Montreal, QC, Canada; 22–25. Available from: https://www.researchgate.net/publication/253877269_Reduction_of_Bleedthrough_in_Scanned_Manuscript_Documents
7. Savino P, Tonazzini A, Bedini L. Bleed-through cancellation in non-rigidly misaligned recto–verso archival manuscripts based on local registration. Int J Doc Anal Recognit. 2019;22:163–176. Available from: https://link.springer.com/article/10.1007/s10032-019-00323-2
8. Savino P, Tonazzini A. Training a shallow NN to erase ink seepage in historical manuscripts based on a degradation model. Neural Comput Appl. 2024;36:11743–57. Available from: https://link.springer.com/article/10.1007/s00521-023-09354-7
9. Hu X, Lin H, Li S, Sun B. Global and local features-based classification for bleed-through removal. Sens Imaging. 2016;17:9. Available from: https://link.springer.com/article/10.1007/s11220-016-0134-7
10. Ettari A, Brescia M, Conte S, Momtaz Y, Russo G. Minimizing Bleed-Through Effect in Medieval Manuscripts with Machine Learning and Robust Statistics. J Imaging. 2025;11:136. Available from: https://doi.org/10.3390/jimaging11050136
11. Seaward L, Kallio M. Transkribus: Handwritten Text Recognition technology for historical documents. In: Workshop Digital Humanities. 2017. Available from: https://dh2017.adho.org/abstracts/649/649.pdf
12. Momtaz Y, Laccetti L, Russo G. Modular Pipeline for Text Recognition in Early Printed Books Using Kraken and ByT5. Electronics. 2025;14:3083. Available from: https://doi.org/10.3390/electronics14153083
13. Reul C, Christ D, Hartelt A, Balbach N, Wehner M, Springmann U, et al. OCR4all—An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings. Appl Sci. 2019;9(22):4853. Available from: https://doi.org/10.3390/app9224853
14. Xue L, Barua A, Constant N, Al-Rfou R, Narang S, Kale M, et al. ByT5: Towards a Token-Free Future with Pre-Trained Byte-to-Byte Models. arXiv. 2022; arXiv:2105.13626. Available from: https://aclanthology.org/2022.tacl-1.17/