Content representation and analysis: the Magic Project and the Illuminated Dante Project integrated systems for multimedia information retrieval

Stefania Conte[1,*], Gennaro Ferrante[2,†], Lorenza Laccetti[1,†], Andrea Mazzucchi[2,†], Yahya Momtaz[1,†] and Augusto Tortora[1,†]

[1] *Department of Physics "Ettore Pancini," University of Naples Federico II, Italy*
[2] *Department of Humanities, University of Naples Federico II, Italy*

Abstract

From the collaboration between the Department of Humanities and the Department of Physics "Ettore Pancini" of the University of Naples "Federico II", with the addition of three private companies, the "MAGIC" project started in mid-2023. The project aims to create a Service Center for the treatment, conservation, digitization, preservation and enhancement of the archival and book heritage. Starting from the Illuminated Dante project, that is, the digitization (managed by the Department of Humanities) of the illuminated manuscripts of Dante Alighieri's Divine Comedy, dated between the 14th and 15th centuries and preserved in national and international museums, libraries and archives, the MAGIC project as a whole proceeds towards information systems and information retrieval systems that aim to acquire information, to archive it and to preserve it in specific databases, without neglecting its distribution and communication. The analysis of the contents and the extraction of the information also lead to a second objective: the easy readability of the text. Artificial intelligence algorithms are being used as image filtering techniques in what is commonly called the bleed-through effect, where ink bleeds between the front and back of manuscripts, making them difficult to read.

Keywords
Manuscripts, artificial intelligence, web portals, information retrieval

1. Introduction

In May 2023, thanks to the cooperation between the Department of Humanities and the Department of Physics “Ettore Pancini” of the University of Naples “Federico II”, the MAGIC project was launched, a multidisciplinary project with heterogeneous and transversal skills, capable of combining scientific research in the field and the growth of knowledge with the

IIR 2024: 14th Italian Information Retrieval Workshop, September 05–06, 2024, Udine, Italy

† These authors contributed equally.

stefania.conte@unina.it (S. Conte); gennaro.ferrante@unina.it (G. Ferrante); lorenza.laccetti@unina.it (L. Laccetti); andrea.mazzucchi@unina.it (A. Mazzucchi); yahya.momtaz@unina.it (Y. Momtaz); augusto.tortora@unina.it (A. Tortora)

0009-0005-9527-3451 (S. Conte); 0000-0003-4968-9630 (G. Ferrante); 0000-0002-0531-75040 (A. Mazzucchi); 0000-0002-0596-906X (A. Tortora)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

use of innovative technological solutions, first and foremost the digitalization planned for the Illuminated Dante project. The Center intends to experiment and adopt new technologies in the service of information retrieval, in the presence of libraries and archives rich in heritages to be communicated and a public attentive and predisposed to interaction and information sharing. The MAGIC project also stems from the co-partnership with three major industrial companies, leaders in the field of Information and Communications Technology applied to cultural heritage and digitization; companies that have an in-house R&D division and have acquired the necessary know-how to guide all stages of the process, from the creation of information to its archiving and dissemination [1] [2].

2. Multimodality and multimedia in Dante Alighieri’s Commedia

Conventions, incentives and scientific partnership agreements with Italian and foreign curators have supported the digitization and dissemination of a core of 283 illuminated manuscripts of Dante Alighieri's Divine Comedy.

The dedicated web portal, IDP-Illuminated Dante Project (www.dante.unina.it), curated by the Department of Humanistic Studies, displays high-definition digitized images in the JPEG 2000 multi-resolution format of the first 101 manuscripts with copyright concession and using the Mirador viewer [3] [4]. It is a large hypertextual system in which access to the data occurs through open access navigation and exploration by the user, be they a specialist and scholar, or an avid reader of the world of Dante.

The web portal, object of interest, represents a system of textual and multimedia information retrieval, since the information to be managed and searched are in different forms: text and images. The system manages hundreds of textual contents, first of all those concerning the poem of the Divine Comedy, according to the edition edited by Giorgio Petrocchi and divided into the three canticles of Inferno, Purgatory and Paradise, in turn divided into verses. Furthermore, the IDP-Illuminated Dante Project portal reports for each manuscript its own descriptive sheet, considering the bibliographic object under the paleographic, codicological, philological and historical aspects, as well as under the iconographic one, arriving at a very analytical expository analysis, which goes as far as the examination of the individual images that accompany and embellish the code. A search for "Subjects" and "Decorative types" from the boxes on the Home page or directly from the Navigation bar allows the system user to perform an exhaustive and complete search of the text/image relationship. For each image inserted in the portal, the typology, execution technique, subject and macro-subject, keywords, text-image relationship, relationship with the Dante and extra-Dante tradition, and any notes are outlined. The same occurs with the iconographic information that characterizes the "Subjects" field, an alphabetical list that lists protagonists, contexts, things and places present in Dante's Comedy. The description cards follow the Manus OnLine (MOL) platform, and will then be exported via XML TEI-P5, on the portal www.dante.unina.it. This procedure allows the indexer to proceed with the compilation of the initial grids in pre-established or free fields [5] [6] [7].

3. Artificial intelligence and data access

The second goal of the Magic project is the creation of new cultural assets available for the community to enjoy. The primary goal is not only the mere faithful reproduction of manuscripts, but to ensure their easy readability for users, transcending sociocultural boundaries. In fact, an important phase of the project concerns the image quality control system, with the aim of ensuring the efficient readability of the entire information content of the original manuscripts, especially for non-experts, since manuscripts commonly have irregular and difficult-to-read writing.

Ancient manuscripts can be subject to different types of degradation. A frequent problem is the oxidation of inks: their composition causes oxidation, or the production of acidic substances, which gradually penetrate between the recto and verso of the pages and vice versa; this effect is known as bleed through. The approach introduced by the Magic project examines artificial intelligence for content analysis and information extraction. The experimental phase allows us to identify the most suitable and effective techniques to counteract this phenomenon compatibly with its presence and intensity. From the tested approaches, a chosen image filtering technique is based on the enhancement and propagation of spatial coherence structures.

Once the text lines within the image are identified, a combination of spatial and semantic concatenation logic can be applied. This method employs a comprehensive pipeline, utilizing various color spaces and techniques such as contrast enhancement and white border inpainting to improve segmentation accuracy of different document regions. In fact, attenuation occurs in two sequential phases, an identification step, which requires the use of unsupervised machine learning algorithms; and a removal step, also called inpainting, where pixels labeled as bleedthrough are replaced with pixels that mimic the background hue. The proposed method integrates a clustering mechanism paired with Gaussian Mixture Models to distinctively identify and separate foreground, background, and bleed-through elements. Furthermore, these techniques allow to know the entire apparatus of comments and glosses that are rich in the illuminated manuscripts of the different editions of Dante's Comedy, also useful for the purposes of research and philological study of the texts (see Figure 1) [8] [9].

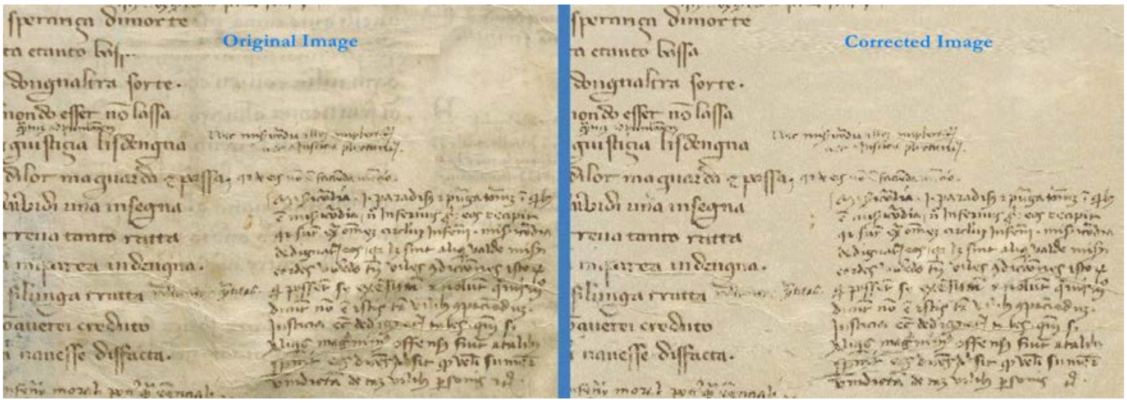

Figure 1: example of bleed through (Copyright: www.dante.unina.it)

4. Conclusions

The interdisciplinarity of the MAGIC project collimates and converges toward the interdisciplinarity of information retrieval, which arises precisely from the intersection of different disciplines involving ontology, design, information architecture, human behavior about information, linguistics, information science, and computer science.

Acknowledgements

The project was funded by Ministry of Enterprise and Made in Italy, (code n. F/130093/03/X38 and CUP: B69J23000560005) and by Ministry of University and Research (code n. PIR01_00011, CUP: I66C18000100006).

References

[1] G. Russo, L. Aiosa, G. Alfano, A. Chianese, F. Cornevilli, G. D. Domenico, P. Maddalena, A. Mazzucchi, C. Muraglia, F. Russillo, A. Salvi, B. Spisso, G. Trombetti, G. Zollo, “MA.G.I.C.: Manuscripts of Girolamini In Cloud” in: IOP Conference Series, Materials Science and Engineering, 949, 012081 (2020), pp. 1–8. doi:10.1088/1757-899X/949/1/012081

[2] S. Conte, P. M. Maddalena, A. Mazzucchi, L. Merola, G. Russo, G. Trombetti in use: 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, Alberto and Fanini, Bruno and Graf, Holger and Pescarin, Sofia and Rizvic, Selma (Eds), Eurographics Workshop on Graphics and Cultural Heritage, The Eurographics Association, 2023, pp. 119-128. doi: 10.2312/gch.20231167

[3] C. Perna, ”Illuminated Dante Project: un archivio e database per la più antica iconografia dantesca (secc. XIV-XV)”, in: DigItalia, 15(2), (2020), pp. 150-158. https://doi.org/10.36181/digitalia-00022.

[4] G. Ferrante , ”Illuminated Dante project : un approccio integrato di studi testuali, librari e iconografici” in: Perna, Ciro (Ed) Immaginare la Commedia, Roma, Salerno, 2022, pp. 237-244.

[5] G. Wan, L. Zi, “Content based information retrieval and digital libraries”, in: Information technology and libraries, (2008), pp. 41-46. doi: 10.6017/ital.v27i1.3262

[6] T. Brizio, “Project Management and Digital Transformation. Performance Measuring Model of Digital Projects and Archives”, in: JLIS.It 9 (3), (2018), pp. 92-108. doi.org/10.4403/jlis.it-12420.

[7] A. Zanni, “Libraries and Open Access Industry”, in: JLIS.It 9 (3), (2018), pp. 75-91. doi.org/10.4403/jlis.it-12486.

[8] S. Conte, G. M. Di Domenico, A. Mazzei, A. Mazzucchi, G. Russo, A. Salvi, A. Tortora, “The MAGIC project: first research results” in: Bernasconi, Eleonora and Mannocci, Andrea and Poggi, Antonella and Salatino, Angelo and Silvello Gianmaria (Eds), Proceedings of the 20th Conference on Information and Research science Connecting to Digital and Library science, Bressanone, Brixen, 22-23 february 2024, 2024, pp. 87-93.

[9] R. Morriello, “Blockchain, intelligenza artificiale e internet delle cose in biblioteca in: AIB Studi, 59 (1-2), (2020), pp. 45-68