The MAGIC project: first research results

Stefania Conte[1], Gian Marco Di Domenico[2], Andrea Mazzei[2], Andrea Mazzucchi[3], Guido Russo[1], Alessandro Salvi[4] and Augusto Tortora[1]

[1] University of Naples "Federico II", Department of Physics "Ettore Pancini", via Cinthia, 21, Naples, 80126, Italy
[2] SA Documents S.r.l., via Seconda della Francesca s.n.c., Cancello ed Arnone, 81030, Italy
[3] University of Naples "Federico II", Department of Humanities, via Porta di Massa, 1, Naples, 80133, Italy
[4] Netcom Group S.p.a., via Nuova Poggioreale, Naples, 80143, Italy

Abstract

This contribution directs its focus towards the initial research findings introduced as part of the MAGIC project. This project emerges from the collaborative efforts between the Department of Humanistic Studies and the "Ettore Pancini" Department of Physics at the University of Naples "Federico II," aimed at establishing a Service Center dedicated to the processing, digitization, preservation, and enrichment of documentary and bibliographic heritage. In elucidating the comprehensive objectives of implementation, we delve into the initial experiments conducted, specifically focusing on the digitization of illuminated manuscripts featuring Dante Alighieri's Divine Comedy.

Keywords

Digitization, long term data preservation, bibliographic resource, Digital preservation, Dante Alighieri,

1. Introduction

In June 2023, the MAGIC project inaugurated a Service Center dedicated to the application of technologies in the realm of cultural heritage. Specifically geared towards the processing of manuscripts, documents, and ancient printed texts, the center's overarching goals encompass the safeguarding and conservation of cultural heritage, alongside initiatives to enhance and improve the readability of these invaluable artifacts.

The synergistic blend of industrial and scientific expertise at the University of Naples Federico II forms the basis for advancing cutting-edge technologies in the sector. This development encompasses the incorporation of Internet of Things (IoT) technologies to safeguard volumes, the utilization of Information Technology for narrative-based usage, the integration of Artificial Intelligence for handwriting and printed character recognition, as well as pre-cataloguing of volumes, and the application of Big Data for seamless access to digitized materials.

The objective is to develop a prototype for a smart library, poised to catalyze the emergence of novel professional standards while simultaneously spearheading the revitalization of the local territory and the advancement of cultural tourism. This holds particular significance in a country like Italy, distinguished by the abundance of its historical, artistic, archaeological, demo-ethno-anthropological, archival, and bibliographic treasures.

The Center endeavors to create culture and foster value through tangible initiatives in restoration, digitalization, and cataloging. Additionally, it strives to impart training to empower young individuals with diverse methodologies and skills, drawing inspiration from the expertise of the proponents and implementers of the project MAGIC.

IRCDL 2024: 20th conference on Information and Research science Connecting to Digital and Library science, February 22–23, 2024, Bressanone-Brixen, Italy

stefania.conte@unina.it (S. Conte); gm.didomenico@consorziocsa.it (G. M. Di Domenico); a.mazzei@consorziocsa.it (A. Mazzei); andrea.mazzucchi@unina.it (A. Mazzucchi); guido.russo@unina.it (G. Russo); a.salvi@netcomgroup.eu (A. Salvi); augusto.tortora@unina.it (A. Tortora)

0009-0005-9527-3451 (S. Conte); 0000-0002-0531-75040 (A. Mazzucchi); 0000-0001-5823-4393 (G. Russo); 0000-0002-0596-906X (A. Tortora)

2. The objectives

The project includes five main directives:

1. *Conservation.* The first directive consists in the very high-resolution digitization of ancient and modern parchments, manuscripts and printed volumes which can be considered as unique pieces due to their material and paleographic peculiarities and due the presence of handwritten annotations and miniatures. The high definition of the reproduction not only ensures the faithful preservation of the original material object but also shields it from external contact, thereby preventing consequential wear and tear. Furthermore, digital acquisition in TiFF and JPG formats is followed by their conversion into F.I.T.S. format. (Flexible Image Transport System), via the back-end interface. The F.I.T.S. format, designed by NASA in the 70s and also used by the Vatican Apostolic Library, is an open standard whose main objective is long-term archiving, transmission and simplification of the use of the data thus saved, according to watchword "Once F.I.T.S., always F.I.T.S.". This implies that data saved in this format will remain accessible and compatible, enduring through any future evolution of the standard.
2. *Access.* The second directive pertains to the accessibility of this cultural heritage, emphasizing the principle of diversifying approaches. From an educational, tourist, and historical-cultural research standpoint, the digital library facilitates the reading of texts and the study of their content across a variety of media, including PCs, tablets, and smartphones. Additionally, the aim is to encourage a dynamic and innovative utilization of the cultural institution and its collection by offering an immersive visit. Visitors can employ Augmented Reality (AR) and Virtual Reality (VR) applications through mobile devices like smartphones or tablets, enhancing their experience with the assistance of a virtual guide.
3. *Preservation.* The third directive concerns the design and testing of sensors, capable of monitoring the state of the asset and carrying out its protection from a safety point of view.
4. *Metadata creation* and the analysis of library assets are the focus of the fourth directive. This directive addresses the extraction and generation of access indexes to facilitate metadata creation and semi-automatic content production, providing a wealth of data to enhance search capabilities. Technological advancement is complemented by a comprehensive campaign to acquire knowledge about each book object. A detailed sheet will be meticulously prepared for each, encompassing its content, structure, cultural history, and preservation status. Following a meticulous analysis of widely adopted standards, the digitization process is accompanied by the creation of an extensive metadata repository. This encompasses descriptions of the manuscript, along with collating existing studies on the subject, creating an interdisciplinary nexus. This elevates the digitized document to a genuine entry point for historical, sociological, and philological exploration, fostering a complete, multilingual, and multicultural journey. It's not just mass digitization of books but rather an operation designed for the masses.
5. *Enhancement*. The fifth directive concerns the use and enhancement of digital heritage at multiple levels:
   a. School. A hyperlink provides access to a brief film that elucidates the book object, showcasing its content and history by featuring a selection of particularly beautiful and/or significant cards/pages;
   b. Tourism. Utilizing an app, you can delve into a concise history of the cultural institute and its primary collection. This includes exploring the interconnectedness between them and understanding their relationship to the city's history;
   c. Culture. Each book is intricately linked to history through various complex facets, encompassing the history of book manufacturing and publishing, the influence of ruling classes who often commissioned these books, the evolution of artistic styles, the history of literary imagery, and the broader historical context of the city.

The activities to be carried out to achieve the final objective of the Project are divided into 8 Implementation Objectives (OR).

• ***OR1***: executive design of a hardware and software architecture for the digitization of texts and subsequent processing. This is followed by the definition of models for image and text processing in order to produce metadata and physical, chemical and biological analyzes of the supports. We adopted the METS standard, which is considered the metadation coding standard. For each volume, we are acquiring two types of metadata: descriptive metadata, focused on describing and facilitating resource search, and structural metadata, aimed at highlighting the relationships between them. The two types have been identified by the National Digital Library Project of the U.S. Library of Congress as being relevant to digital collections. An example of the first kind of metadata is in the dimension parameters (size, aspect ratio), in the collation (e.g. 4 bifoglio - *quaternion*) and so on. An example of the second kind of metadata is in bibliographic references, in manuscript history and so on.
• ***OR2***: Commencing with digitization, an additional endeavor involves implementing an artificial intelligence system proficient in recognizing diverse writing styles, printed and handwritten characters. The designed system falls within the realm of weakly supervised learning, allowing it to enhance its accuracy by leveraging cataloging information;
• ***OR3***: Experimental development of the image acquisition system and the data archiving system described in OR1.
• ***OR4***: Identification and creation of metadata and cataloging parameters for the library material.
• ***OR5***: Creation of a prototype of a conservation, public use, interface and monitoring system adequate for the amount of data to be managed.
• ***OR6***: Creation of knowledge production and Open Data management models and technological analysis for use through augmented reality. The possibility of representing the semantic relationships between the metadata created through Linked Open Data (LOD) technology, and through ontological languages (RDFS) and graph technologies (RDF) is also envisaged. We then proceed by structuring the metadata of the manuscript catalog starting from the Cultural Internet and the international Europeana, Gallica, etc.
• ***OR7***: Generating image masters for long-term preservation employing the F.I.T.S. format, alongside compressed images, tailored for practical use and easy consultation.
• ***OR8***: User portal development, along with the configuration of the immersive visit app.

3. The various phases of the project

Operationally, the MAGIC project is divided into 3 phases, which involve various cultural institutes in Naples, as well as different aspects in which a book can take shape:

1. ***Illuminated manuscript codices of Dante Alighieri's Divine Comedy***. This is a nucleus of approximately 280 manuscripts dating back to between the 14th and 15th centuries and preserved in national and international libraries, museums, public and private archives (for example the National Vittorio Emanuele III Library of Naples, the Civic Historical Archive and the Trivulziana Library of Milan, the Central National Library of Florence, the Vatican Apostolic Library, the National Library of France, section manuscripts, the National Library of Spain, the Morgan Library and Museum of New York).
2. ***Incunabula and sixteenth century works belonging to the Pontaniana Academy of Naples***. In the second phase we proceed with the digitization of the 15th and 16th century editions, preserved in the library of the Accademia Pontaniana, an academy based in Naples, founded around 1443. In particular, these are 6 incunabula and 186 sixteenth century works, coming from the collection of Francesco and Luigi Torraca which, following the latter's death, were donated by his heirs to the Academy. A single collection that brings together works of history, law, philosophy, Greek, Latin, Jewish and Italian literature, religion, law, architecture, medicine, numismatics, astronomy, geography, magic and astronomy.
3. ***Manuscripts from the Girolamini Library of Naples***. The third phase involves the digitization of selected incunabula belonging to one of the richest libraries, the Girolamini Library of Naples which includes various book collections, from which works of literature,

philosophy, Christian theology, philosophy, history of the Church and music come sacred. For this phase, an agreement with the library is being defined.

Digitization is carried out at high resolution using planetary scanners and according to the guidelines of the International Federation of Library Associations and Institutions ("Guidelines for Planning the Digitization of Rare Book and Manuscript Collections, 2015").

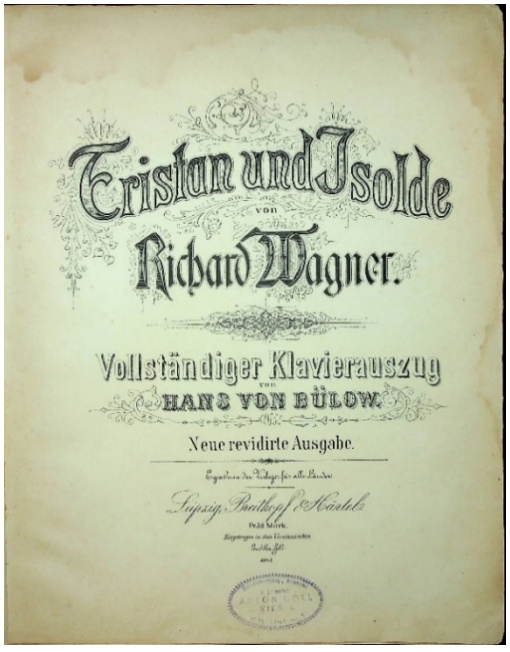
4. The participating subject

The MAGIC project involves the collaboration with both public and private partners. Among the private entities, namely Netcom Engineering S.p.a., SA Lombardia S.r.l., and SA Documents S.r.l. as the leading partner, specialize in providing expert consultancy in alignment with CNIPA and Ministry of Culture regulations. For the project implementation, these companies leverage cutting-edge technologies in information processing, including textual databases, hypertext, and image processing, contributing to the enhancement and utilization of cultural heritage.

The University of Naples “Federico II” is present in the team as a Research Organization (OdR). The project engages the expertise of both the Department of Humanistic Studies and the "Ettore Pancini" Department of Physics, each contributing their specialized skills. The former is dedicated to education and research in humanistic disciplines, encompassing areas such as archaeology, cultural heritage, philology, philosophy, literature, languages, psychology, cultural heritage sciences, history, and the history of the arts. The latter department oversees and coordinates research activities in Universe Physics, Nucleus and Radiation Physics, Subnuclear and Astroparticle Physics, Theoretical Physics, Applied Physical Methodologies, the Structure of Matter, and Informatics within its realm of expertise.

5. Prototyping phase

Prior to initiating any digitization endeavors, an experimentation phase was conducted to accommodate the diverse forms that bibliographic material may assume. The inaugural prototype focused on a printed musical score from the latter half of the 19th century, a part of the Mathematics Department's collection at the University of Naples. Specifically, the chosen material was Richard Wagner's "Tristan und Isolde," distinguished by the inclusion of handwritten annotations made by Renato Caccioppoli, a renowned Italian mathematician and music enthusiast.



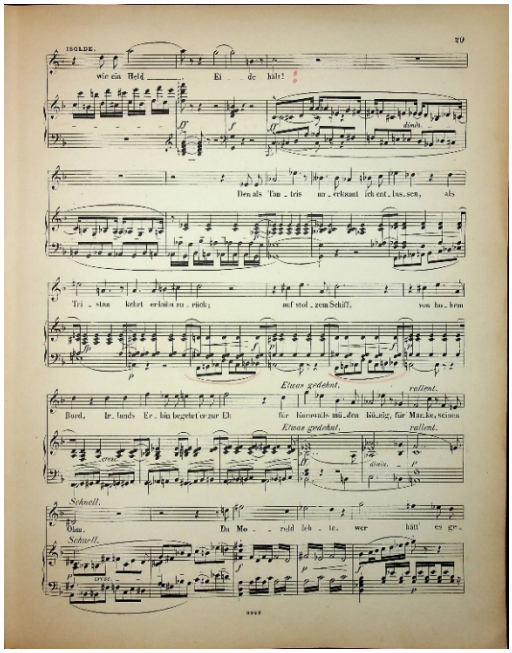


Figure 1: Musical score “Tristan und Isolde” by Richard Wagner (Copyright: department of Mathematics, University of Naples)

A second bibliographic typology concerned the printed monograph, coming from the collection of the Pontaniana Academy, "Memorie della regale Accademia ercolanese di archeologia" of 1862.

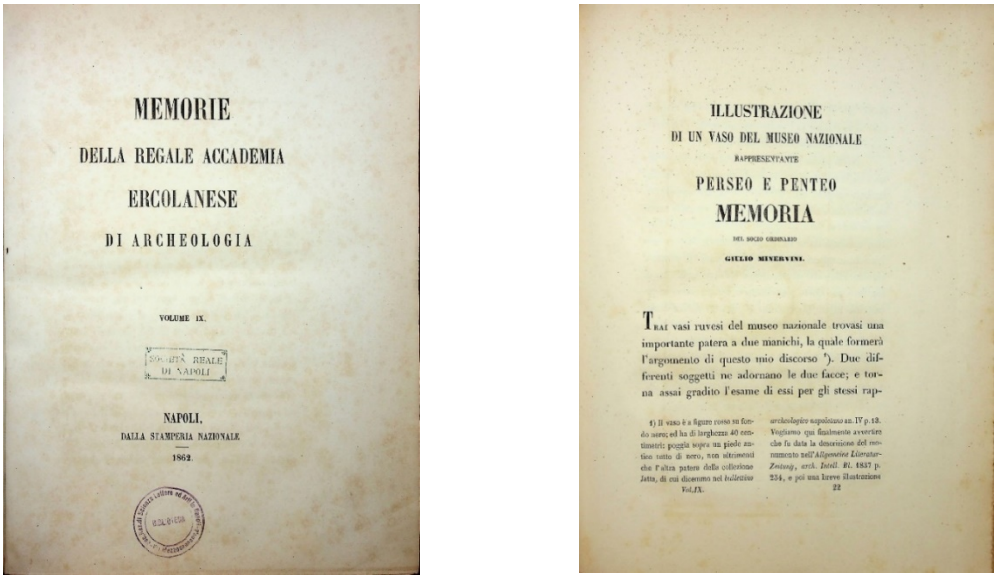


Figure 2: Printed volume “Memorie della regale Accademia ercolanese di archeologia” (Copyright Accademia Pontaniana)

With the help of a planetary scanner, high quality jpeg files were prepared with 24 bit color depth and 330 dpi optical resolution. The conditions of both volumes influenced the digital acquisition process and the image quality control system. This latter is now based on color calibration according to the ISO 12641-2:2019 standard, followed by a visual inspection, but we are developing an automated method to verify both colors and geometry according to predefined metrics. Furthermore, the prototype phase has included the conversion of the files into the F.I.T.S. format.

6. First research results

Through a thematic digitization initiative that has yielded 93 illuminated manuscripts of Dante Alighieri's Divine Comedy, sourced from libraries, museums, and national and international archives, a valuable virtual collection has been curated. As of December 2023, this collection is accessible (*https://www.dante.unina.it/public/frontend)*, enriched with comprehensive codicological and iconographic information. The primary aim is not mere faithful reproduction of the manuscripts but rather ensuring their easy readability for users, transcending socio-cultural boundaries. Emphasizing a spirit of sharing and an open access model, the objective is continual expansion. The codices, dating from the 14th to the 15th centuries, come with detailed descriptive elements, transforming into high-quality metadata. This metadata encompasses the manuscript's history and tradition, meticulously covering aspects such as binding, collation, incipits and explicit text, with accompanying images spotlighting the illuminated manuscripts.

An equally crucial phase involves implementing an image quality control system, with the goal of guaranteeing the efficient readability of all the informational content within the original manuscripts. A fundamental process for ensuring optimal visualization, especially for non-experts, is the application of image filtering techniques. Manuscripts commonly exhibit irregular writing, where text lines are not consistently horizontal or straight, and their angles may vary. Identifying and rectifying such irregularities in text lines are essential steps in achieving high-

quality character recognition. This ensures not only good visualization but also facilitates easy reading for a broader audience.

From the tested approaches, a chosen image filtering technique relies on the enhancement and propagation of spatial coherence structures. Once the lines of text within the image are identified, a combination of spatial and semantic concatenation logic can be applied.

Another common challenge, often encountered in manuscript papers, is the oxidation of inks, particularly those of iron-gallic composition. This oxidation results in the production of acidic substances that gradually seep into the papers, potentially causing perforation. In the realm of research, various approaches in artificial vision are being explored to identify the most effective techniques to visually mitigate this effect.

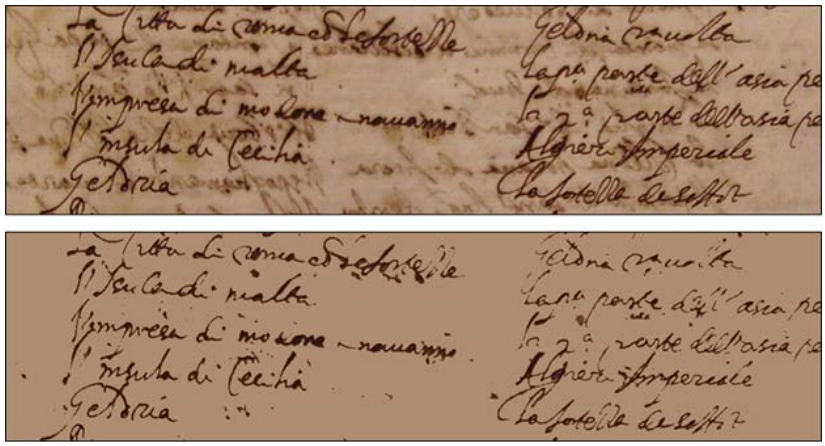


Figure 3: Effect reduction approach bleed through

All Dante's codes, objects of interest, are characterized by ornamental or illustrative elements, miniatures, friezes, diagrams which have an intrinsic artistic value. It follows that the extraction and comparative analysis of these elements is extremely important for artistic and historical research purposes. The extraction of the ornamental contents of the manuscripts is formulated, once again, as a semantic image segmentation problem, using appropriate convolutional neural architectures.

The semantic segmentation neural network for segmenting ornaments in ancient manuscripts stands as a robust solution owing to its inherent capacity for intricate pattern recognition. Leveraging advanced deep learning techniques, this model excels in discerning subtle details and intricate designs within the manuscript, achieving high accuracy by precisely delineating the boundaries of ornamental elements. The network's ability to analyze contextual relationships and subtle variations in ornamentation contributes to its efficacy, ensuring a nuanced and accurate segmentation process that significantly outperforms traditional methods. This sophisticated approach represents a cutting-edge advancement in the realm of manuscript analysis, providing a reliable and precise means for extracting meaningful information from intricate visual data.

The MAGIC laboratory has pioneered a novel approach for the economy dissemination of knowledge—empowering users to navigate formats, organizational structures, and languages previously unfamiliar to them. This initiative aims to expand their skills, fostering a future demand for knowledge. Aligned with the vision of widespread utilization, the MAGIC project endeavors to create diverse opportunities: making valuable information available to a select group of users while concurrently introducing it to a new audience. This involves adapting traditional forms of use and repurposing them for alternative contexts such as social media, innovative services, and storytelling formats. The MAGIC laboratory has developed software that facilitate the creation of the Flipping Page effect, enabling the rotation of pages through digitized content accessed via tablets or smartphones. Essentially, knowledge enhancement is a collaborative effort involving both specialized and non-specialized users. The preliminary

resampling operation ensures that scanned images undergo further processing, including the removal of margins without compromising the aspect ratio, followed by standardizing the number of pixels along both the x and y axes.

7. Conclusions

The MAGIC project is in its nascent stage, marking the commencement of a substantive inquiry. The nature of this endeavor necessitates access to archaic manuscripts, a privilege contingent upon judicious authorization, a process characterized by its inherent complexity and occasional protraction. Of equal significance is the imperative for a systematic user evaluation, wherein feedback from prototypical use-case scenarios becomes instrumental. Rigorous scrutiny of accessibility, usability, and the lucidity of informational content will be conducted by not only adept researchers but also a broader cross-section of readers. This initiation signals the commencement of a scholarly odyssey, wherein the forthcoming chapters promise to unveil a tapestry of progressive advancements.

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] L. Andreoli, M. Cimino, G. Drago, Linee Guida sulla digitalizzazione di documenti bidimensionali, (2022) Università degli Studi di Padova - Sistema Bibliotecario di Ateneo

[4] F. Mercanti, " Controller digital lending e digital lending: un confronto sulla lettura digitale tramite le biblioteche" in AIB Studi, 63 (2), (2023), pp. 279-295

[5] Linee guida per pianificare la digitalizzazione di collezioni di libri rari e manoscritti, International Federation of Library Associations and Institutions, in: A. Nuovo, C. Cauzzi, C. Consonni, V. De Francesca, D. Deana, L. Longhi, F. Viazzi (Eds), 2015

[6] 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

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

[8] 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.

[9] S. Allegrezza, "Analisi del formato FITS per la conservazione a lungo termine dei manoscritti. Il caso significativo del progetto della Biblioteca Apostolica Vaticana" in: DigItalia, 6(2), (2011), pp. 43–72

[10] M. Lana, "Digital humanities e biblioteche" in AIB Studi, 59 (1-2), (2020), pp. 185-223

[11] F. Niutta, "Manoscritti nella rete" in: DigItalia, 5(2), (2010), pp.9-28