Looking back to 1850 in 2025: HistoraScan to digitize historical journals

Bruno Frutuoso Costa

CIES-Iscte, ISCTE - University Institute of Lisbon

Address: Avenida das Forças Armadas, nº40, 1649-026 Lisboa, Portugal

Email: bruno_frutuoso_costa@hotmail.com

ORCID ID: https://orcid.org/0000-0003-3023-8960

This manuscript is the accepted version of the following book chapter:

Costa, B. F., Mateus, B. C., Pinto, H. J., & Tabrizi, M. R. (2025). Looking Back to 1850 in 2025: Historascan to Digitize Historical Journals. In A. Rosário & A. Boechat (Eds.), *Impact of Digitalization on Communication Dynamics* (pp. 393-420). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3693-3579-6.ch015

Looking back to 1850 in 2025: HistoraScan to digitize historical journals

Bruno Frutuoso Costa
CIES-Iscte, Iscte - University Institute of Lisbon, Portugal

Bruno Contreiras Mateus
IADE – European University of Lisbon and Iscte - University Institute of Lisbon, Portugal

Hugo José Pinto
Inovaworks II Command and Control S.A, Portugal

Mohammad Reza Tabrizi
Inovaworks II Command and Control S.A., Portugal

ABSTRACT
This chapter analyses current technologies and the challenges involved in extracting and classifying articles and news headlines from historical journals, as well as converting images to text format. The work to develop a tool focused on digitising historical journals entitled "HistoraScan" was carried out by a multidisciplinary team of experts in media studies, artificial intelligence, image processing, and cultural heritage preservation. The data used derives from two historic Portuguese journals, Diário de Notícias and Jornal de Notícias, which were created in the mid-19th century. This project is based on a mixture of heuristics, computer vision, pattern recognition, and other artificial intelligence and machine learning techniques. The main challenges included the variability in the design of historical journals, preserving the quality of images over time, and continuously improving image processing and OCR techniques to adapt to different styles and periods of newspapers.

Keywords: Industry 4.0, News media, Historical newspapers, Journals, Digitization, Innovation, Content creation, Artificial intelligence, Machine learning, Optical Character Recognition, Convolutional Neural Networks, Natural language processing, Historascan, Diário de Notícias, Jornal de Notícias.

INTRODUCTION

The Fourth Industrial Revolution and digital transformation are expressions used to designate the paradigm in which cutting-edge technologies are used in automation and data exchange to help increase the quality, efficiency, and productivity of companies' production processes (Gilchrist, 2016; Skilton & Hovsepian, 2018). However, Industry 4.0 is not limited to the digitization of equipment, procedures, data collection, and IoT networks (Barros et al., 2023a, 2023b). Based on the principles of interconnectivity, information transparency, decentralized decisions, and technical assistance (Hermann et al., 2016), technological developments have made scalable cloud and service solutions available, such as software as a service (SaaS), platform as a service (PaaS), and hybrid cloud. Their combination with cyber-physical systems, algorithms (which integrate implicit business logic), and cybersecurity mechanisms in a mixed infrastructure allocated to public and private datacentres allows for the development of integrated information management solutions. In a complementary sense, artificial intelligence provides these infrastructures with the ability to learn and produce precise, complex, and sophisticated forecasts that support decision-making and the definition of new business directions (Iansiti & Lakhani, 2020; Skilton & Hovsepian, 2018).

The false belief that Industry 4.0 is synonymous with complex and expensive tools only available to multinational companies has prevented many companies from investing in the digitization of key operations in their value chain and in research and development (R&D) capabilities for value-added products and services to respond to market challenges and opportunities (Gilchrist, 2016). In the creative industries, innovation emerged in the UK at the end of the 20th century, encompassing sectors that combine creativity and communication technologies. These include news media, performing arts, visual arts, music, film and video, design, advertising, and leisure software, to name a few (Granado et al., 2020). The main objective was to create economic and social value for

companies. At the beginning of the 21st century, innovation was closely linked to the emergence of the Internet (Bonixe, 2020).

Based on the assumption that digital transformation demands speed, breadth, and depth of change from companies (Schwab, 2017; Takeuchi & Nonoka, 2008/2004), the automotive, technological, and biological industries took on the capacity to innovate and lead this paradigm shift early on (Coelho, 2016). In journalism, companies have been forced to migrate to digital platforms at the risk of jeopardizing their revenues, causing various challenges in terms of routines, languages, ethics, and training for professionals (Bonixe, 2020). Innovation in the media industry varies significantly between countries and companies, being influenced by financial factors, organizational culture, and government regulation (Costa & Mateus, in press; de-Lima-Santos & Ceron, 2021; Meier et al., 2022; Pérez-Seijo & Vicente, 2022). As a rule, the adaptation of the business model of newspaper companies occurs slowly and with few forms of innovation (Cardoso et al., 2019; Crespo et al., 2018).

Among the most common innovations are the use of digital platforms and social networks (Meier et al., 2022), the creation of new multimedia formats (Bonixe, 2020; Gehlen & Sousa, 2018; Granado et al., 2020; Silva & Granado, 2021; Vicente & Pérez-Seijo, 2022), the development of immersive apps (Fante, 2018; García-Avilés et al., 2019), the implementation of data analysis tools (de-Lima-Santos & Ceron, 2021), and the adoption of digital business models (Cardoso et al., 2019). In the area of artificial intelligence, data analysis to identify trends and patterns, automatic fact-checking, and the creation of subtitles and automatic transcripts for videos are little used in the news industry (de-Lima-Santos & Ceron, 2021), except for countries such as the USA and China. In these two countries, artificial intelligence has been used to automate repetitive tasks, such as report writing, and to personalize relevant content for news websites. These actions are supported by large data sets that reflect users' interests and behaviours in the digital environment and with companies (de-Lima-Santos & Ceron, 2021).

Innovation and its integration are a complex process that requires a combination of resources, collaborative leadership, and a predisposed organizational culture (García-Avilés et al., 2019; Meier et al., 2022; Pérez-Seijo & Vicente, 2022). Many organizations have limited resources to invest in

expensive innovations, which can hinder the adoption of new technologies and infrastructures (García-Avilés et al., 2019; Meier et al., 2022; Pérez-Seijo & Vicente, 2022). Journalists and other professionals may be more resistant to using new technologies because they feel they detract from their core mission of reporting events or because of the perception that they could threaten their work, but also because they lack the technical skills to implement them effectively (Costa & Mateus, in press; García-Avilés et al., 2019; Meier et al., 2022; Pérez-Seijo & Vicente, 2022). Meier et al. (2022) point out that sustainable business models are scarce. Several innovations have proven not to be economically viable in the long term, which can make their adoption difficult and companies fearful of them.

In this area, the automatic digitization of historical newspapers (and other similar publications) has been the subject of research due to the enormous amount of information they contain, and the challenges associated with the variability in the presentation of documents over time, such as the configuration and pagination of the support, as well as the language used, which have evolved over relevant periods (Kumpulainen & Late, 2022; Lemmers et al., 2023; Rhyno, 2017; Ringel, 2023).

The aim of this chapter is to analyse current technologies and the challenges involved in extracting and classifying articles and news headlines from historical newspapers and converting historical newspaper images to text format, in order to present the work carried out to develop a tool focused on digitizing historical newspapers: The HistoraScan.

BACKGROUND

The automatic digitization of historical journals represents a growing field of research of great importance for the preservation and accessibility of cultural heritage (Allen & Hall, 2010; Powell & Paynter, 2009; Skelbye & Dannells, 2021). This area of study focuses on the development of technologies and methods to convert physical collections of old journals into digital formats, thus enabling easier consultation and more efficient preservation of the documents (Kettunen et al., 2020; Koistinen et al., 2017). The aim is not only to create digital copies but also to extract textual and visual information in an accurate and useful way,

facilitating research and public access to these valuable historical sources (Ehrmannm et al., 2022).

In recent years, several approaches have been proposed to address the technical challenges of automatically digitising historical journals. Work such as that by Rhyno (2017) highlights the use of advanced optical character recognition (OCR) techniques to deal with the typographic variability and imperfections of old documents. In addition, projects such as the Library of Congress' “Newspaper Navigator”[1] use convolutional neural networks (CNNs) to improve the accuracy of extracting images and text from digitised journals, highlighting the importance of deep learning techniques in this area.

The integration of artificial intelligence (AI) tools is beginning to be studied because of its potential to revolutionise the digitization of historical journals (Spina, 2023; Teel, 2024; Ferro, et al., 2023), with particular attention to OCR error correction and document segmentation (Dhali, 2024). Interdisciplinary collaboration between computer scientists, historians and archivists has been fundamental to progress in this area (Thomas et al., 2024). Collaborative projects such as ‘READ’[2] (Recognition and Enrichment of Archival Documents) exemplify how the application of pattern recognition and AI technologies can be enhanced by historical and archival expertise to develop more robust solutions adapted to the specific needs of historical archives. These initiatives not only improve the accuracy of digitisation, but also promote the democratisation of access to historical heritage, allowing a wider audience to explore and value these rich sources of knowledge.

Automatic digitization of historical newspapers

The automatic digitization of historical newspapers refers to the automated process of converting physical historical newspapers into a digital format readable by computers (Allen et al., 2008). This process is essential for preserving and making accessible the information contained in old newspapers, which are valuable resources for historical and cultural research (Bunout et al., 2023). It facilitates access to and exploration of vast collections of historical newspapers, allowing researchers to conduct analyses and studies in a more

[1] https://labs.loc.gov/work/experiments/newspaper-navigator/
[2] https://www.transkribus.org/

efficient and comprehensive manner. Additionally, automatic digitization helps preserve the information contained in these documents, protecting it from potential physical deterioration or loss over time

The automatic digitization of historical journals has several applications that cover multiple areas of study and interest. As outlined in Table 1, these applications include research in journalism and media studies, where academics can analyse digitised journals to study the evolution of journalistic practices, media representations, and communication trends over time. In the field of historical research, historians and researchers have the possibility of accessing digitised journals to study past events, trends, and cultural phenomena, allowing specific searches by keywords, topics, or dates. In the area of genealogy, individuals can use digitised journals to trace family history, finding information such as birth announcements, obituaries, wedding notices, and other family events. Cultural heritage institutions such as libraries and archives can use digitization to preserve and make accessible their collections, ensuring that valuable cultural materials are protected for future generations and shared with a wider audience. Furthermore, in the field of education, teachers and students can use digitised historical journals as educational resources, analysing primary sources, examining historical perspectives, and participating in activities that promote critical thinking and research skills.

Table 1. Applications of automatic digitization of historical newspapers

Areas	Examples
Journalism and media studies	Media scholars can analyse digitized newspapers to study the evolution of journalism practices, media representations, and communication trends over time. They can explore changes in reporting styles, editorial policies, and societal attitudes reflected in historical news coverage.
Historical research	Scholars and historians can access digitized historical newspapers to

	study past events, trends, and a cultural phenomenon. They can search for specific keywords, topics, or dates to gather information and insights.
Genealogy	Individuals interested in tracing their family history can use digitized newspapers to find information about their ancestors, such as birth announcements, obituaries, marriage notices, and other family-related events.
Cultural heritage preservation	Libraries, archives, and cultural institutions can digitize historical newspapers to preserve and make accessible their collections. This ensures that valuable cultural heritage materials are safeguarded for future generations and can be shared with a wider audience.
Education	Teachers and students can use digitized historical newspapers as educational resources. They can analyse primary sources, examine historical perspectives, and engage in activities that promote critical thinking and research skills.

Source: Own elaboration

The process involves the use of technologies such as Optical Character Recognition (OCR), image processing, artificial intelligence, and machine learning (Kettunen et al., 2022; Fizaine et al., 2024; Valente et al., 2023). They

allow to perform tasks such as converting images of newspaper pages into editable digital text, segmenting content (headlines, articles, images, among others), improving image quality, and organizing digitized data into a suitable format.

The automatic digitization of historical newspapers has been a subject of research for years due to the vast amount of information they contain, and the challenges associated with the variability in document presentation over time (Kumpulainen & Late, 2022; Lemmers et al. 2023; Rhyno, 2017; Ringel, 2023). The challenges are essentially associated with document structure analysis, segmentation and labelling, image processing and OCR, Convolutional Neural Networks (CNN) and Natural Language Processing (NLP), data storage and format, and integration of headlines and article.

A document image is made up of a variety of physical regions. Examples: blocks of text, lines, words, figures, tables, and backgrounds. According to Namboodiri and Jain (2007), the process of analysing the structure and layout of a document aims to break down a given document image into its regions and components, understanding their functional roles and relationships. If we consider that in historical journals, the main elements are sentences in columns, headings, titles, and captions, as well as images and author names, the main challenge is to assign functional labels to these elements (Girdhar et al., 2024).

Segmenting historical journals into different classes (headlines, articles, images, and advertising, among others) involves appropriate image segmentation and labelling techniques. This requires using computer vision to improve image quality and segmentation for enhanced OCR processing. The last one is used to convert text images into editable digital text (Palfray et al., 2012).

CNN is a multi-layer neural network method that learns the hierarchical characteristics of data and is commonly used for image classification tasks in this context (Wang & Gang, 2018). In recent years, CNN has evolved rapidly in the design and calculation of NLP. In turn, NLP can be used to improve OCR accuracy and help extract semantic information from journalistic texts. It allows computers to work on word segmentation (WB), information extraction (IE), relation extraction (RE), named entity recognition (NER), part-of-speech tagging (POS), word sense disambiguation (WSD), text-to-speech (TTS), machine translation (MT), among others (Wang & Gang, 2018).

As the amount of material to be processed is vast and breaks down into various levels of detail, the metadata needs to be reliable to facilitate automatic text processing. It is responsible for guiding OCR, linguistic processing and search (Allen & Schalow, 1999). When storing and formatting data, it is important to organise the digitised data in an appropriate format, dividing the data set into training, test and validation sets and adjusting the model's hyperparameters to obtain the best results. In the case of metadata, these provide reference points that determine the performance of programmes. However, it can also later be used by end users as a guide for searching and interacting with collections of historical journals.

Grouping headlines with their respective articles is a crucial step in facilitating access to information and improving the usability of digitised data (Liebl & Burghard, 2020; Schultze et al., 2024). The quality of the machine-readable text on a journal page can be improved when individual blocks of text are identified as such before OCR is carried out (Klijn, 2008). Segmentation techniques are generally a semi-automatic process, so manual checking of the results is often necessary (Broersma & Harbers, 2018). For example, articles whose text appears on different pages of the journal may need to be followed up on manually. However, this is a time-consuming task, so the solution is to improve the consistency of the journal's layout. The more predictable the layout, the better the automated segmentation (Broersma & Harbers, 2018; Klijn, 2008).

All over the world, libraries offer free access to digitised historical journals through user interfaces (Pestana, 2021). In the initial phase, only features such as search and filter options were made available. As users increasingly desire more options and more advanced tools, another challenge emerges, which concerns the development of simple, usable interfaces for different user groups (Pfanzelter et al., 2021).

In this area, companies specialized in the digitization of historical documents can be direct competitors. These companies offer services to convert physical documents into digital formats, including historical newspapers (Ehrmann et al., 2019). Companies focused on OCR and image processing may be indirect competitors. They provide solutions to convert image text into digital text and may have features that address the specific challenges of digitizing historical newspapers (Rhyno, 2017). On the other hand, some software

providers offering document management solutions may be competitors, as some of them may have features related to the organization and classification of historical data (Black, 2023; Schlukbier, 2008).

MAIN FOCUS OF THE CHAPTER

Sample

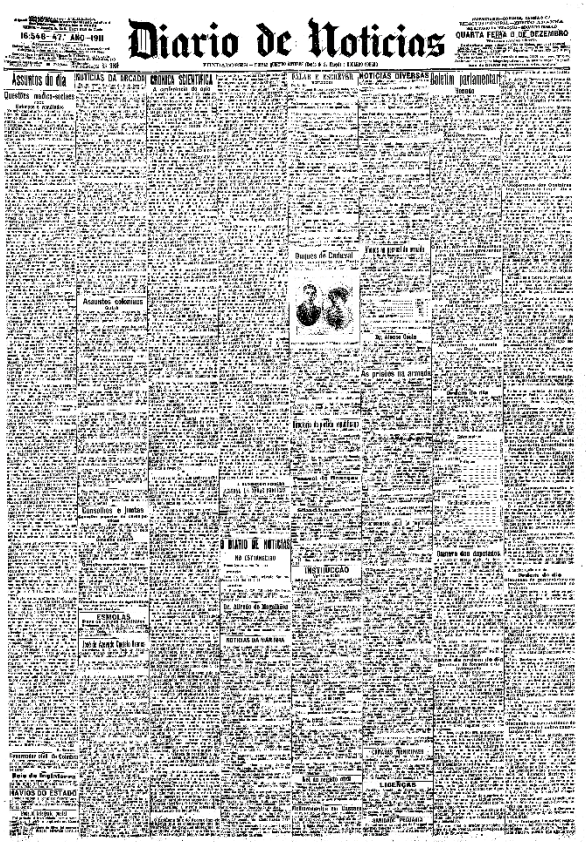
Given the difficulties in obtaining a set of public data that would provide relevant information and allow the development of the HistoraScan tool to digitise historic journals, we decided, after several searches and attempts, to use a set of data provided by Diário de Notícias (Figure 1) and Jornal de Notícias. These journals are considered as reference presses in the sense that their target audience is leading public opinion, and their content focuses on national and international politics, economics, and culture (Costa & Antunes, 2024; Figueiras, 2005). To this end, they essentially use an analytical and explanatory journalistic angle on the subject, standing out from the rest for their ability to be an agenda-setter (Figueiras, 2005). This means that "their highlights also become the priorities of the public agenda (themes shared by the community), of the other media, and also of the capacity to condition the political agenda, imposing behaviour on it" (Figueiras, 2005, p. 4).

A newspaper or journal is a periodical publication that focuses on the dissemination of news, reports, opinions, and announcements of general or specific interest, usually on a daily or weekly basis. Newspapers are aimed at a wide audience and aim to inform readers about recent events, current affairs, and developments in various areas, including politics, economics, sports, and entertainment. Typical examples include The New York Times and The Guardian. Their accessibility and comprehensive coverage of topics make newspapers a primary source of daily news for the public (Franklin, 2016; McNair, 2017). Newspapers are produced in short time cycles, with tight deadlines to ensure that the news is current, and involve a team of editors, reporters, and photographers working to cover day-to-day events. In addition, journals generally have a wide and diverse target audience (Sterling, 2009).

Despite several attempts to publish cheap newspapers in Portugal, it was with the arrival of the Industrial Revolution that Diário de Notícias (DN) was

founded in 1864 by journalist and businessman Eduardo Coelho and his partner and printer Tomás Quintino (Sousa, 2018, 2021). At the time, the doctrinaire, combative, and partisan press characterised the Portuguese journalistic field. This was categorised by a high degree of rotation between the main parties in government. The opposition newspapers fought the government party journals, and vice versa. In this way, DN's industrial model presented itself as politically independent, giving rise to a newsroom with several dozen professionals, in which the reporter became increasingly important. Writing techniques quickly became oriented towards the pursuit of factuality, the separation between information and opinion, and the objectivity, as can be seen in this passage: "It was a journal different from the other Portuguese newspapers of the time, in content (news), in style (clear, concise, precise, and simple), in form, namely in appearance (four-column layout, not two or even one, as was customary)" (Sousa, 2018, p. 171).

Figure 1. Historical image from the Portuguese newspaper "Diário de Notícias"


Diario de Noticias

Source: Own material

Jornal de Notícias (JN) was founded in 1888 by José Diogo Arroio, Manuel Vaz de Miranda, and Aníbal da Costa Morais. Unlike DN, the founding members belonged to the Regeneration Party (Helena, 2017). In this way, the monarchist journal often argued against the government's progressivism, publishing national and international news with a page of commercials. In 1907, it was the journal with the largest circulation in the north of the country, but in that year, it abandoned its monarchist editorial line and adopted a republican slant in defence of the interests of the north. Although with the inclusion of a somewhat sensationalist narrative. The changes implemented enshrined a popular press style that was accessible to a wider audience than the one traditionally made up of political, social, and intellectual elites, the journal's audience par excellence (Helena, 2017).

Currently, Portuguese journalism faces significant challenges, such as the pressure on journalists to produce content quickly. Lack of resources is the main barrier to innovation in the industry (Cardoso et al., 2019). Crespo et al. (2018) show that journalistic companies in Portugal are slowly adapting to new practices. They favour the use of different available narratives and tend to edit with new publishing tools, these being the main forms of innovation. It is considered that the serious economic crisis that journalism has faced in the last decade has reduced teams that don't have the time or technical skills to develop and test new forms of innovation (Bonixe, 2020; Fante, 2018; Miranda et al., 2021; Silva & Granado, 2021; Vicente & Pérez-Seijo, 2022).

Methods

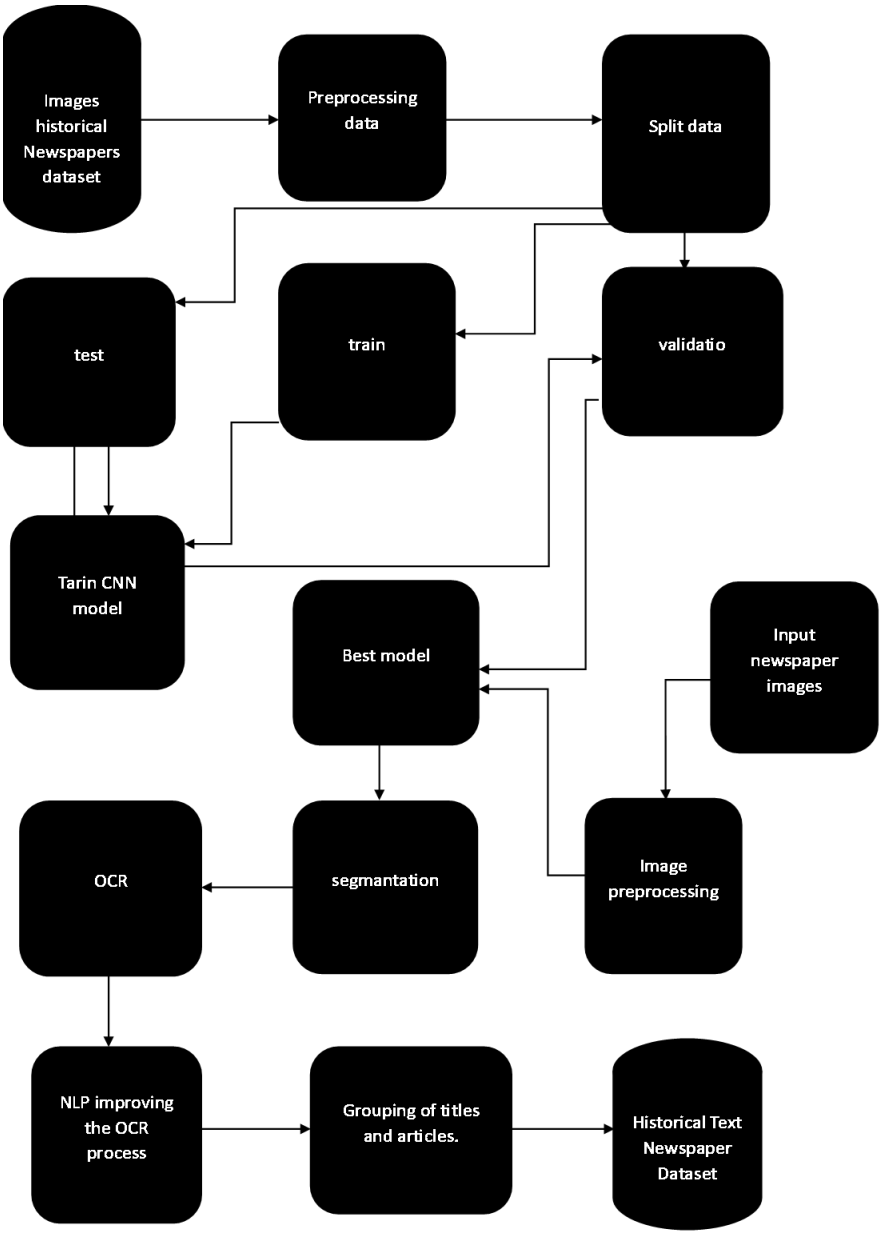
The DN and JN dataset provides data from the last century to the present day. These images contain information about headlines, articles, advertisements, among others. To understand the data, we were going to work with, an exploratory analysis was carried out in the first phase.

The work was carried out by a multidisciplinary team of experts in media studies, artificial intelligence, image processing, and cultural heritage preservation. This project involved a series of new techniques to adapt and evolve the state of the art in retrieving data from newspaper pages published

since 1850, based on a mixture of heuristics, computer vision, pattern recognition, and other artificial intelligence and machine learning techniques.

As Figure 2 illustrates, a sequence of page scanning, image segmentation, natural language processing, and heuristics was used to reconstruct first a single article segment, then a complete article, and finally a page and complete relationships between pages. Work to refine these techniques with pilot tests in an iterative process is also discussed.

Figure 2. Steps used in the development of HistoraScan



Source: Own elaboration

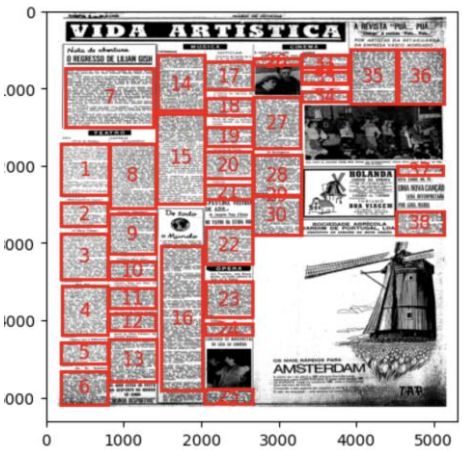
Models developed for segmentation

In our tool, we utilized CNNs for segmenting pages of historical newspapers into distinct sections such as articles, titles, page headers, images, and advertisements (Figure 3 and Figure 4). By leveraging CNNs, which are particularly adept at processing visual data, we aimed to accurately delineate the various components of newspaper pages.

Through extensive training and optimization of the CNN models, we achieved precise segmentation results, enabling efficient extraction of content from digitized newspaper pages. This segmentation process not only facilitates the organization of newspaper data but also enhances the accessibility and usability of historical archives for researchers and enthusiasts.

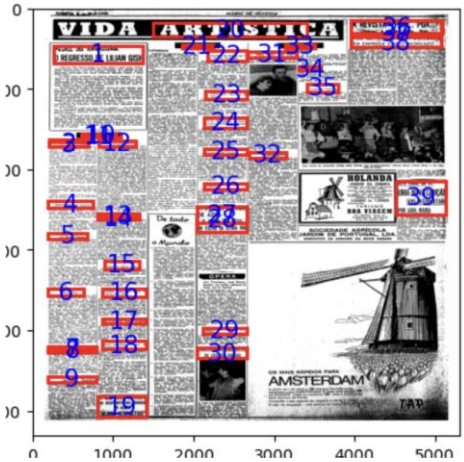
Additionally, by automating the segmentation task with CNNs, we streamlined the digitization process, saving time and resources compared to manual segmentation methods. The results demonstrate the effectiveness of CNN-based approaches in segmenting complex document layouts and paves the way for further advancements in historical newspaper digitization and analysis. As we can see in Figures 3 and 4, the segmentation of the newspaper image occurred with the identification of several filters that allowed, in this case, only the title and article to be selected.

Figure 3. Segmentation result of articles



Source: Own elaboration

Figure 4. Segmentation result of titles



Source: Own elaboration

OCR

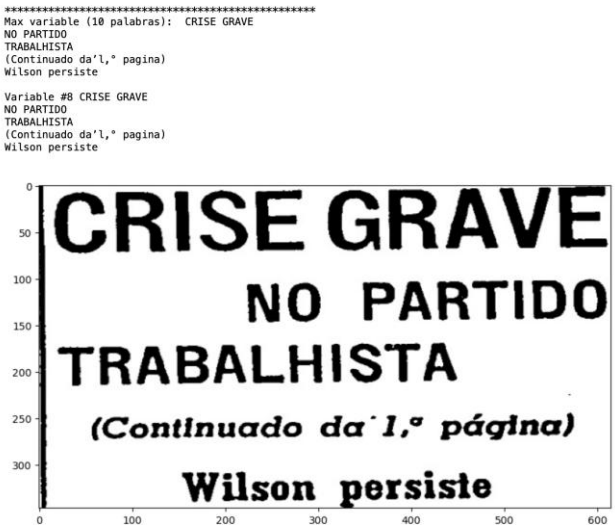
After segmenting images, we utilize image segmentation algorithms using computer vision techniques to enhance image qualities, focusing on specific regions such as titles, articles, headers, and images. This segmentation process involves identifying and isolating different components within the image to improve accuracy and readability (Figure 5 and Figure 6). Subsequently, OCR is employed to convert these segmented images into text format. OCR algorithms analyse the content of each segmented region, recognizing characters and patterns to generate a textual representation of the image content.

This involves preprocessing steps to enhance image clarity and remove noise, followed by the application of OCR algorithms to extract text from the segmented regions. Finally, the extracted text is processed and refined using natural language processing techniques to improve accuracy and coherence, ensuring that the converted text retains the original meaning and context from the segmented images. Overall, this process integrates image segmentation, OCR, and natural language processing to effectively convert segmented images into readable text data.

Figure 5. OCR process (Phase 1)

Source: Own elaboration

Figure 6. OCR process (Phase 2)



Source: Own elaboration

Both figures showcase the extracted text from historical documents, enabling digitization and subsequent manipulation of the text for analysis, search, or any other necessary purpose.

NLP models

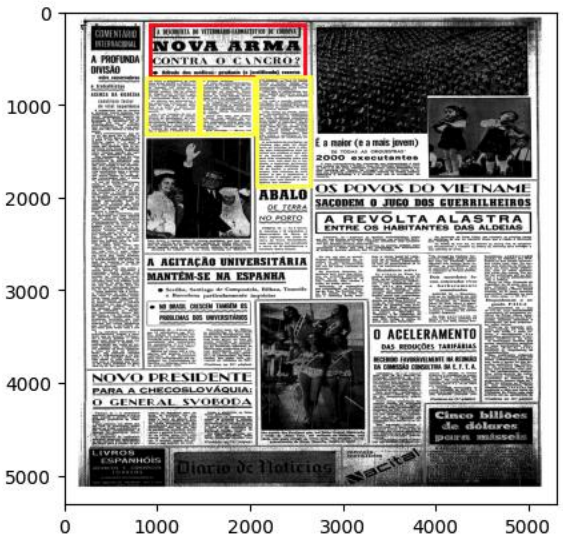
After the OCR process, some outputs may exhibit errors in their original spelling or word separation, particularly due to the complexity of historical newspaper images, which often feature faded text and irregular layouts. To

address this challenge and enhance the accuracy of our tool, we employ NLP networks to improve the results in this regard.

Correlating headlines and articles

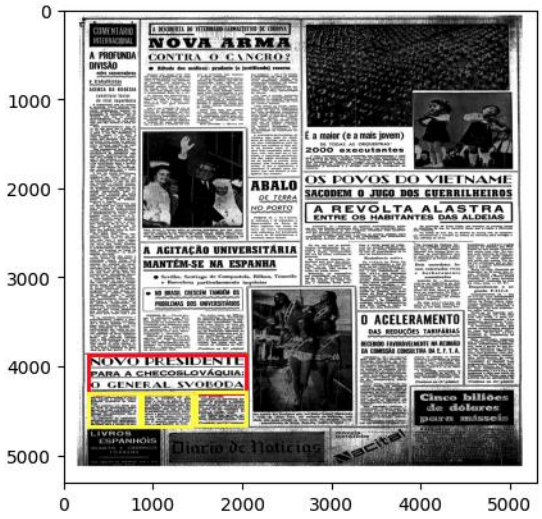
We propose employing logic and various algorithms to mitigate discrepancies between newspaper titles and articles, aiming for a more precise understanding of their relationship. This involves conducting an in-depth analysis of the structure and semantic content of each title and its corresponding article (Figure 7 and Figure 8). NLP algorithms can be employed to discern linguistic patterns and common themes between titles and texts. Additionally, data mining techniques may unveil significant associations based on keyword frequency and content similarity. The utilization of machine learning algorithms, such as classification and clustering, could aid in organizing titles and articles into coherent thematic groups. This approach promises to deepen comprehension of the interplay between newspaper elements, thereby facilitating access to pertinent information for diverse research and analytical purposes.

Figure 7. Correlating headlines and articles



Source: Own elaboration

Figure 8. Correlating headlines and articles



Source: Own elaboration

To finalize the process, we save the extracted text from both the headlines and articles into separate text files in TXT format. This allows for easy access and further analysis of the textual content. This step ensures that the valuable information extracted through the OCR process is preserved and can be utilized for various purposes such as text mining, sentiment analysis, and information retrieval. Developing this idea further, saving the extracted text in TXT format enables researchers and analysts to conduct in-depth textual analysis and explore patterns, themes, and insights within the historical newspaper data. By having the headlines and articles stored in separate files, it facilitates efficient data management and enables targeted analysis of specific sections of the newspapers. Additionally, the TXT format is widely supported and can be easily imported into various text processing tools and platforms for further exploration and research purposes.

DISCUSSION

The proposed methodology effectively addressed the challenges associated with historical newspaper digitization and text extraction. By employing a combination of computer vision techniques, OCR processes, and

NLP models, we successfully segmented images, extracted textual content, and correlated headlines with articles. The integration of advanced technologies such as neural networks and natural language processing significantly improved the accuracy and efficiency of the data processing pipeline.

Furthermore, the developed models and algorithms showcased promising results in accurately extracting and correlating textual information from historical newspapers. The ability to store the extracted text in TXT format facilitates further analysis and research, enabling scholars to delve deeper into the historical content and uncover valuable insights.

INNOVATIONS BROUGHT BY HISTORASCAN TO THE MARKET

Our tool specifically focuses on the unique challenges associated with digitizing historical newspapers, which sets it apart from generic document digitization solutions. By utilizing advanced techniques such as CNN and NLP, HistoraScan can offer greater accuracy and efficiency in extracting and classifying information from historical newspapers.

Given that the design of historical newspapers may vary over time, HistoraScan may include specific features to adapt to these changes and ensure accurate and consistent digitization. On the other hand, this tool provides an easy-to-use interface and efficient processes for digitization and data organization.

CONCLUSION

As previously demonstrated, HistoraScan focuses on employing neural networks for the segmentation of images from historical newspapers, utilizing various computer vision techniques to enhance the quality of the segmented images. Additionally, we employed an OCR process to convert these images into text format, subsequently leveraging NLP techniques to refine the accuracy of the OCR output.

The main challenges included the variability in the design of historic newspapers, preserving the quality of images over time, and continually improving image processing and OCR techniques to adapt to different newspaper styles and periods. In this sense, HistoraScan has the potential to be the most innovative platform on the market for digitizing, classifying, and searching for

information in historical newspaper collections. This tool is aimed at researchers, historians, and history enthusiasts around the world.

The digitization of historical newspapers can offer numerous advantages to contemporary print journals, promoting a rich exchange between the past and the present. Firstly, making digitised archives available allows contemporary journalists to access a wide range of historical references, enriching the contextualization and depth of current stories. This can result in more informed and comprehensive journalism, where reports not only cover present events, but also draw parallels and lessons from the past, offering a more complete perspective to readers. In addition, analysing historical journalistic practices can inspire innovations in editorial approaches and storytelling, taking advantage of techniques that were effective in other times and adapting them to contemporary needs and expectations.

Another significant added value is the potential revitalization of public interest and reader loyalty. By integrating historical content with current news, print journals can attract a wider audience, including historians, researchers, educators, and history enthusiasts. Subscription programmes or dedicated history sections within journals can increase the value perceived by readers, encouraging deeper engagement with the published content. Additionally, digitization allows journals to tell stories about their own evolution, creating an institutional narrative that strengthens the newspaper's identity and brand. This not only preserves institutional memory, but also celebrates the journal's role in society over time, establishing a stronger and more emotional connection with its audience.

Overall, this study contributes to the field of digital humanities by providing a comprehensive framework for digitizing and analysing historical newspapers. The insights gained from this research have the potential to enhance historical research, facilitate cultural heritage preservation, and stimulate interdisciplinary collaborations in the digital humanities domain.

ACKNOWLEDGMENT

This research was supported by the Portuguese Foundation for Science and Technology (FCT – Fundação para a Ciência e a Tecnologia) [2023.04877.BD].

REFERENCES

Allen, R. B., & Schalow, J. (1999). Metadata and data structures for the historical newspaper digital library. In *Proceedings of the eighth international conference on Information and knowledge management (CIKM '99)* (pp. 147–153). Association for Computing Machinery.

Allen, R. B., Waldstein, I., & Zhu, W. (2008). Automated processing of digitized historical newspapers: Identification of segments and genres. In G. Buchanan, M. Masoodian, S. J. Cunningham (Eds.), *Digital libraries: Universal and ubiquitous access to information* (pp. 379–386). Springer.

Allen, R. B., Hall, C. (2010). Automated Processing of Digitized Historical Newspapers beyond the Article Level: Sections and Regular Features. In G. Chowdhury, C. Koo, & J. Hunter (Eds.), *The Role of Digital Libraries in a Time of Global Change.* Springer.

Barros, C., Costa, B. F., & Ramos, D. (2023a). *Guia de boas práticas para a indústria 4.0* [Best practices guide for industry 4.0]. UA Editora.

Barros, C., Costa, B. F., & Ramos, D. (2023b). I*ndustry 4.0: Best practices for the digital transition*. UA Editora – Universidade de Aveiro.

Black, J. W. (2023). Creating specialized corpora from digitized historical newspaper archives: An iterative bootstrapping approach. *Digital Scholarship in the Humanities*, *38*(2), 779–797.

Bonixe, L. (2020). Jornalismo radiofónico e inovação: Uma análise à cobertura de acontecimentos mediáticos [Radio journalism and innovation: An analysis of coverage of media events]. *Media & Jornalismo*, *20*(36), 153-169.

Broersma, M., & Harbers, F. (2018). Exploring machine learning to study the long-term transformation of news: Digital newspaper archives, journalism history, and algorithmic transparency. *Digital Journalism*, *6*(9), 1150–1164.

Bunout, E., Ehrmann, M., & Clavert, F. (2023). *Digitised newspapers: A new eldorado for historians? Reflections on tools, methods and epistemology*. De Gruyter Oldenbourg.

Cardoso, G., Baldi, V., Crespo, M., Pinto-Martinho, A., Pais, P. C., Paisana, M., & Couraceiro, P. (2019). *O que devem saber os jornalistas? Práticas e formação em Portugal* [What should journalists know? Practices and training in Portugal]. OberCom.

Coelho, P. M. (2016). *Rumo à Indústria 4.0* [Towards Industry 4.0] [Unpublished master dissertation]. University of Coimbra, Coimbra, Portugal.

Costa, B. F., & Antunes, E. (2024). Dinâmicas sociais no Facebook: Análise de comentários em conteúdos jornalísticos sobre a suposta tentativa de ataque terrorista à FCUL [Social dynamics on Facebook: Analysis of comments on journalistic content about the alleged attempted terrorist attack on FCUL]. *Observatorio (OBS*)*, *18*(1), 126-150.

Costa, B. F., & Mateus, B. C. (in press). Innovating in journalism with newsgames: An exploratory study in Portugal. In L. Bojić, S. Žikić, J. Matthes, & D. Trilling (Eds.), *Navigating the digital age: An in-depth exploration into the intersection of modern technologies and societal transformation*. Institute for Philosophy and Social Theory.

Crespo, M., Foà, C., & Pinto-Martinho, A. (2018). Como o jornalismo lida com a inovação: Um estudo de caso das melhores práticas em Portugal [How journalism deals with innovation: A case study of best practices in Portugal]. *Estudos de Jornalismo*, 9, 75-102.

de-Lima-Santos, M.-F., & Ceron, W. (2021). Artificial Intelligence in news media: Current perceptions and future outlook. *Journalism and Media*, *3*(1), 13–26.

Dhali, M. A. (2024). *Artificial Intelligence in Historical Document Analysis: Pattern recognition and machine learning techniques in the study of ancient manuscripts with a focus on the Dead Sea Scrolls* [Unpublished doctoral dissertation]. University of Groningen, Groningen, Netherlands.

Ehrmann, M., Bunout, E., & Düring. (2019, August). *Historical newspaper user interfaces: A review* [Paper presentation]. 85th IFLA General Conference and Assembly (IFLA), Athens, Greece.

Ehrmannm M., Düring, M., Neudecker, C., & Doucet, A. (2022). Computational Approaches to Digitised Historical Newspapers. *Dagstuhl Reports*, *12*(7), 112-129.

Fante, A. (2018). Jornalismo appificado: Arquiteturas da notícia [Appified journalism: News architectures]. *Estudos de Jornalismo*, 9, 112-124.

Ferro, S., Pelillo, M., & Traviglia, A. (2023). AI-assisted digitalisation of historical documents. In *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci*. (pp. 557–562). XLVIII-M-2-2023.

Figueiras, R. (2005). Os comentadores na imprensa de referência portuguesa: 1980-1999 [Commentators in the Portuguese reference press: 1980-1999]. In *Anais Intercom - XXVIII Congresso Brasileiro de Ciências da Comunicação* (pp. 1-15). Rio de Janeiro.

Fizaine, F. C., Bard, P., Paindavoine, M., Robin, C., Bouyé, E., Lefèvre, R., & Vinter, A. (2024). Historical text line segmentation using deep learning algorithms: Mask-RCNN against U-Net Networks. *Journal of Imaging* ,*10*(3), 1-18.

Franklin, B. (2016). *The Future of Journalism: In an Age of Digital Media and Economic Uncertainty*. Routledge.

García-Avilés, J. A., Carvajal-Prieto, M., Arias, F., & De Lara-González, A. (2019). How journalists innovate in the newsroom. Proposing a model of the diffusion of innovations in media outlets. *The Journal of Media Innovations*, *5*(1), 1–16.

Gehlen, M. A., & Sousa, J. P. (2018). Jornalismo de dados em Portugal: Um estudo exploratório sobre práticas jornalísticas especializadas [Data journalism in Portugal: An exploratory study on specialized journalistic practices]. *Estudos de Jornalismo*, 9, 125-139.

Gilchrist, A. (2016). *Industry 4.0: The Industrial Internet of Things*. Apress.

Girdhar, N., Coustaty, M., & Doucet, A. (2024). Digitizing history: Transitioning historical paper documents to digital content for information retrieval and

mining: A comprehensive survey. In *IEEE Transactions on Computational Social Systems* (vol. 11, pp. 6151-6180).

Granado, A., Silva, D. S., & Vicente, P. N. (2020). Inovação nos media e nas indústrias criativas limítrofes – uma introdução [Innovation in the media and neighboring creative industries – an introduction]. *Media & Jornalismo*, *20*(36), 5-9.

Helena, L. (2017). A evolução editorial do Jornal de Notícias e a inclusão de elementos de jornalismo popular [The editorial evolution of Jornal de Notícias and the inclusion of elements of popular journalism]. *Revista Portuguesa de História da Comunicação*, 1, 1-11.

Hermann, M., Pentek, T. & Otto, B. (2016). Design Principles for Industrie 4.0 Scenarios. In *Proceedings of 49th Hawaii International Conference on System Sciences (HICSS)* (pp. 3928-3937). USA.

Iansiti, M., & Lakhani, K. R. (2020). *Competing in the age of AI: Strategy and leadership when algorithms and networks run the world*. Harvard Business Review Press.

Kettunen, K., Koistinen, M., & Kervinen, J. (2020). Ground Truth OCR Sample Data of Finnish Historical Newspapers and Journals in Data Improvement Validation of a re-OCRing Process. *Liber Quarterly*, 30, 1-20.

Kettunen, K., Keskustalo, H., Kumpulainen, S., Pääkkönen, T., & Rautiainen, J. (2022, February). *OCR quality affects perceived usefulness of historical newspaper clippings: A user study* [Paper presentation]. 18th Italian Research Conference on Digital Libraries, Padova, Italy.

Klijn, E. (2008). The current state-of-art in newspaper digitization: A Market Perspective. *D-Lib Magazine*, *14*(1/2).

Koistinen, M., Kettunen, K., & Paakkonen, T. (2017). Improving Optical Character Recognition of Finnish Historical Newspapers with a Combination of Fraktur & Antiqua Models and Image Preprocessing. In *Proceedings of the 21st Nordic Conference of Computational Linguistics* (pp. 277–283). Sweden.

Kumpulainen, S., & Late, E. (2022). Struggling with digitized historical newspapers: Contextual barriers to information interaction in history research activities. *Journal of the Association for Information Science and Technology*, *73*(7), 1012–1024.

Lemmers, F., Ott, M, & Hermans, S. (2023). Printed mass media and automatic digitisation: The case of Belgian illustrated magazines from the interbellum. *Monte Artium*, 15, 1-31.

Liebl, B., & Burghard, M. (2020, November). *From historical newspapers to machine-readable data: The origami ocr pipeline* [Paper presentation]. Workshop on Computational Humanities Research, Amsterdam, Netherlands.

McNair, B. (2017). *An Introduction to Political Communication*. Routledge.

Meier, K., Schützeneder, J., García-Avilés, J. A., Valero-Pastor, J. M., Kaltenbrunner, A., Lugschitz, R., Porlezza, C., Ferri, G., Wyss, V., & Saner, M. (2022). Examining the most relevant journalism innovations: A comparative analysis of five European countries from 2010 to 2020. *Journalism and Media*, *3*(4), 698–714.

Miranda, J., Fidalgo, J., & Martins, P. (2021). Jornalistas em tempo de pandemia: Novas rotinas profissionais, novos desafios éticos [Journalists in times of pandemic: New professional routines, new ethical challenges]. *Comunicação e Sociedade*, 39, 287-307.

Namboodiri, A. M., & Jain, A. K. (2007). Document structure and layout analysis. In B. B. Chaudhuri (Ed.), *Digital document processing: Advances in pattern recognition* (pp. 29-48). Springer.

Palfray, T., Hebert, D., Nicolas, S., Tranouez, P., & Paquet, T. (2012). Logical segmentation for article extraction in digitized old newspapers. In *Proceedings of the 2012 ACM symposium on Document engineering (DocEng '12)* (pp. 129–132). USA.

Pestana, O. (2021). O acesso aos jornais históricos: Considerações sobre o desenvolvimento de coleções digitalizadas [Access to historical newspapers: Considerations for developing digitized collections]. *Media & Jornalismo*, *21*(39), 161-174.

Pérez-Seijo, S., & Vicente, P. N. (2022). After the hype: How hi-tech is reshaping journalism. In J. Vázquez-Herrero, A. Silva-Rodríguez, MC. Negreira-Rey, C. Toural-Bran, X. López-García (Eds.), *Total Journalism. Studies in Big Data* (pp. 41-52). Springer.

Pfanzelter, E., Oberbichler, S., Marjanen, J., Langlais, P., & Hechl, S. (2021). Digital interfaces of historical newspapers: Opportunities, restrictions and recommendations. *Journal of Data Mining & Digital Humanities*, 1-26.

Powell, T., & Paynter, G. (2009). Going Grey? Comparing the OCR Accuracy Levels of Bitonal and Greyscale Images. *D-Lib Magazine*, *15*(3/4).

Rhyno, A. (2017, April). *Historical newspaper digitization on a shoestring* [Paper presentation]. 2017 IFLA International News Media Conference. Landsbókasafn Íslands – Háskólabókasafn, Reykjavík, Iceland.

Ringel, S. (2023). Digitizing the paper of record: Archiving digital newspapers at the New York Times. *Journalism*, *24*(2), 245-261.

Schlukbier, G. (2008). Digitization of old newspapers: Software developments to enable successful conversion. In H. Walravens (Ed.), *Newspapers collection management: Printed and digital challenges* (pp. 143-146). K. G. Saur.

Schultze, C., Kerkfeld, N., Kuebart, K., Weber, P., Wolter, M., & Selgert, F. (2024). Reading yesterday's news. Layout recognition by segmentation of historical newspaper pages.

Schwab, K. (2017). *The fourth industrial revolution*. Crown Publishing Group.

Silva, D. S., & Granado, A. (2021). *Cobertura jornalística dos números da Covid-19: Casos de inovação em Portugal* [Journalistic coverage of Covid-19 numbers: Cases of innovation in Portugal]. Obi.Media.

Skelbye, M. B., & Dannells, D. (2021). OCR Processing of Swedish Historical Newspapers Using Deep Hybrid CNN–LSTM Networks. In *Proceedings of Recent Advances in Natural Language Processing* (pp. 190–198). Online.

Skilton, M., & Hovsepian, F. (2018). *The 4th Industrial Revolution: Responding to the impact of artificial intelligence on business*. Palgrave Macmillan.

Sousa, J. P. (2018). Eduardo Coelho e a fundação do Diário de Notícias [Eduardo Coelho and the founding of Diário de Notícias]. In J. P. Sousa (Ed.), *Notícias em Portugal: estudos sobre a imprensa informativa (século XVI-XX) [News in Portugal: studies on the informative press (16th-20th century)]* (pp. 163-192). Colecção Livros ICNOVA.

Sousa, J. P. (2021). *Portugal. Pequena história de um grande jornalismo I. Da manufatura à indústria* [Portugal. Small story of great journalism I. From manufacturing to industry]. Livros ICNOVA.

Spina, S. (2023). Artificial Intelligence in archival and historical scholarship workflow: HTS and ChatGPT. *Umanistica Digitale*, 16, 125-140.

Sterling, C. H. (2009). *Encyclopedia of Journalism*. Sage Publications.

Takeuchi, H., & Nonoka, I. (2008). *Gestão do conhecimento* [Knowledge management] (A Thorell, Transl.). Bookman. (First published in 2004).

Teel, Z. (2024). Artificial Intelligence's Role in Digitally Preserving Historic Archives. *Preservation, Digital Technology & Culture*, *53*(1), 29-33.

Thomas, A., Gaizauskas, R., & Lu, H. (2024, May). *Leveraging LLMs for Post-OCR Correction of Historical Newspapers* [Paper presentation]. Lingotto Conference Centre, Torino, Italy.

Valente, J., António, J., Mora, C,. & Jardim, S. (2023). Developments in image processing using deep learning and reinforcement learning. *Journal of Imaging*, *10*(9), 1-22.

Vicente, P. N., & Pérez-Seijo, S. (2022). Spatial audio and immersive journalism: Production, narrative design, and sense of presence. *Profesional de la información*, *31*(5), 1-14.

Wang, W., & Gang, J. (2018). Application of Convolutional Neural Network in Natural Language Processing. In *2018 International Conference on Information Systems and Computer Aided Education (ICISCAE)* (pp. 64-70). Jilin Finance Building Hotel.

KEY TERMS AND DEFINITIONS

Automatic digitization of historical journals: an automated process of converting physical historical journals into a digital format readable by computers.

Convolutional Neural Networks (CNN): a multi-layer neural network method that learns the hierarchical characteristics of data and is normally used for image classification tasks.

Fourth industrial revolution: expression used to designate the paradigm in which cutting-edge technologies are used in automation and data exchange to help increase the quality, efficiency, and productivity of companies' production processes.

Innovation: creation of economic and social value for companies.

Journal: A periodical publication dedicated to the dissemination of news, reports, opinions, and announcements of general or specific interest, usually on a daily or weekly basis.

Natural Language Processing (NLP): a sub-field of artificial intelligence that focuses on the interaction between computers and human language. The aim is to enable computers to understand, interpret, process, and generate natural language in a way that is valuable and meaningful.

Optical Character Recognition (OCR): technology that makes it possible to convert different types of documents, such as scanned paper documents, PDF files, or images captured by a digital camera, into editable and searchable data.