# AI-driven Extraction of Structured Data from a Large Newspaper Archive

by Nikos Kontonasios, Yannis Tzitzikas (FORTH-ICS and University of Crete), and Pavlos Fafalios (FORTH-ICS and Technical University of Crete)

***The project PortADa investigates ship arrivals at the ports of Barcelona, Marseille, Havana, and Buenos Aires from 1850 to 1914, aiming to generate an open-access database containing thousands of records related to 19th-century maritime trade. As part of this effort, the Centre for Cultural Informatics of FORTH-ICS is developing a data extraction pipeline for ship arrivals recorded in the historical newspaper Le Sémaphore de Marseille (1827–1944). The pipeline integrates document layout analysis, optical character recognition (OCR), and generative AI-driven techniques to extract relevant data from scanned newspaper images and store it in a well-structured, machine-readable format for further analysis.***

Historical newspaper archives serve as invaluable sources of information, documenting not only political and economic changes but also cultural, social and technological developments [1, 2]. PortADA is a European project (MSCA Staff Exchanges) in the field of maritime history that draws on historical maritime newspapers to study maritime trade, economic history, and port activities in the 19th and early 20th centuries [L1]. It focuses on ship arrivals at the ports of Barcelona, Marseille, Havana and Buenos Aires from 1850 to 1914. The project's objective is to extract ship arrivals data from local newspaper archives and develop an open-access database containing thousands of records that document 19th-century maritime trade.

Bulletin de la Vigie du Commerce

MOUVEMENT DU PORT

ARRIVÉES

DÉPARTS

MANIFESTES D'ENTRÉE

NOUVELLES DE MER

Figure 1. Example of a newspaper section providing information about ship arrivals (source of newspaper image: retronews.fr).

The project is coordinated by the University of Barcelona and involves multiple partners, including the Institute for Mediterranean Studies of FORTH (IMS-FORTH), the Institute of Computer Science of FORTH (ICS-FORTH), the Autonomous University of Madrid, the Maritime Museum of Barcelona, and institutions in Argentina and Cuba. IMS and ICS-FORTH play a key role in extracting ship arrival records for the port of Marseille, using the historical newspaper Le Sémaphore de Marseille as its primary source, which consistently reported all ship arrivals in Marseille in its daily issues. The newspaper's archives span from 19 December 1827 to 19 August 1944, with 35,703 issues published during this period. The full collection is publicly accessible through the Bibliothèque Nationale de France (BNF)'s dedicated newspaper archive website [L2].

Given the scanned images of a newspaper issue, the task is to extract well-structured data located in a specific newspaper section. In the case of the Le Sémaphore de Marseille, the focus is on ship arrivals, which are documented in a section named "ARRIVÉES", as shown in Figure 1. In this section, information about each ship arrival is presented in small paragraphs arranged vertically on the page. A ship arrival paragraph contains the following information: port of origin, departure date, port and date of call, ship type, ship's flag, ship name, ship tonnage, name of captain, cargo information, consignee of cargo, consignee of ship, intelligence (events or interactions during the journey).

Automating the extraction of this data from all newspaper issues presents several interconnected challenges. The first challenge is segmenting complex newspaper layouts into distinct text blocks while preserving the correct reading order. The second challenge involves accurately locating and isolating the ship arrivals section within each newspaper issue. The third challenge is parsing this section into individual paragraphs, each corresponding to a single ship arrival. The fourth challenge is converting the unstructured text of each paragraph into a structured format with specific data points. Finally, the fifth challenge stems from the poor quality of historical newspaper images, which suffer from physical deterioration, fading, smudging, skew, and curvature, all of which can significantly impact text recognition [3].

To address these challenges, we have developed an automated pipeline of integrated solutions that transforms the scanned newspaper images into a structured dataset. The various stages of this process are illustrated in Figure 2 and described in detail below.

Text Block Identification: The first stage establishes the layout and structure of the newspaper page by separating text into individual, clearly defined blocks. The tool used for this task is Arcanum's newspaper segmentation API [L3]. It takes as input an image of a newspaper page and outputs a JSON file containing information about the distinct blocks.

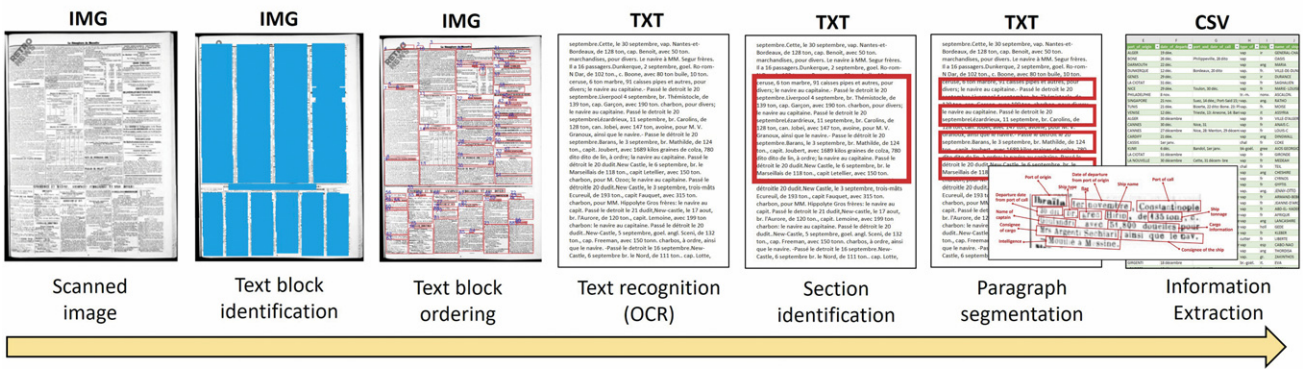


Figure 2. Overview of the information extraction process.

Text Block Ordering: The text blocks identified in the previous step are often not in a sequential reading order due to the layout complexities. To address this, a dedicated software application was developed, which first determines the segments contained in the newspaper page by applying a colour filling to all text blocks using OpenCV [L4]. The blocks are then horizontally stretched so that all blocks within the same segment merge together, effectively highlighting the distinct segments. Then, the columns within each segment are identified using a similar method. Finally, each segment and its respective columns are sorted to ensure a coherent reading order.

Text Recognition (OCR): Once ordered, the text blocks are processed using Google Vision AI to convert the visual text in the images into machine-readable format (text data). Despite its high accuracy, the OCR output can still contain errors, often due to image quality issues.

Section Identification: Because of possible OCR errors, identifying the arrivals section requires more than a simple exact match search for "ARRIVÉES". To address this, the Levenshtein distance method has been applied, which locates the word closest to "ARRIVÉES" as well as a word resembling "DEPARTS" that appears after it and denotes the end of the arrivals section.

Paragraph Segmentation: The segmentation of the ship arrivals section into individual paragraphs is carried out using ChatGPT 4o, leveraging its text processing capabilities to identify paragraph breaks and formatting cues within the text. For this, a carefully designed prompt was employed which directs the AI model to segment the text into paragraphs without modifying the content.

Information Extraction: The distinct data elements for each ship arrival are extracted from the segmented paragraphs by using ChatGPT's natural language processing capabilities. A detailed prompt was used to guide the extraction process, which explicitly defines the fields to be extracted and instructs the AI model to focus on the raw input text without attempting to interpret or correct OCR errors, thereby preserving the authenticity of the data.

Preliminary evaluation results demonstrate high pipeline performance, achieving robust accuracy in text block ordering, acceptable OCR quality with an average accuracy of 82.3%, and high precision in paragraph segmentation (95.7% F1-score). The information extraction stage performed reliably, achieving an average F1-score of 96% across twelve information fields. These results highlight the system's strengths while also identifying areas for future refinement.

A key strength of the pipeline is its modularity, which allows for targeted adjustments and improvements to individual components. This flexibility ensures that the system can be easily adapted to suit other documents or extraction tasks.

This work has received funding from the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101129889 (Project PortADa).

**Links:**

[L1] https://www.proyectoportada.eu/
[L2] https://www.retronews.fr/titre-de-presse/semaphore-de-marseille
[L3] https://www.arcanum.com/en/newspaper-segmentation/
[L4] https://opencv.org/

**References:**

[1] A. Bingham, "The digitization of newspaper archives: Opportunities and challenges for historians," Twentieth Century British History, vol. 21, no. 2, pp. 225–231, 2010.
[2] B. Nicholson, "The digital turn: Exploring the methodological possibilities of digital newspaper archives," Media History, vol. 19, no. 1, pp. 59–73, 2013.
[3] J. Jarlbrink and P. Snickars, "Cultural heritage as digital noise: Nineteenth century newspapers in the digital archive," Journal of Documentation, vol. 73, no. 6, pp. 1228–1243, 2017.

**Please contact:**

Pavlos Fafalios
FORTH-ICS and Technical University of Crete, Greece
fafalios@ics.forth.gr