An Open-Source solution for layout parsing of historical Italian newspapers

Silvano Imboden[1,‡], Antonella Guidazzoli[1,‡], Donatella Sforzini[2,‡], Luca Mattei[2,‡], Gabriele Marconi[1,‡], Federico Andrucci[1,‡], Alex Gianelli[1,‡], Simona Caraceni[1,‡], Rossella Pansini[1,‡], Gabriele Fatigati[1,‡], Fauzia Albertin[1,‡], Giovanni Baisi[1,‡], Giorgia Cardano[1,‡], Maria Chiara Liguori[1,*,‡] and Angela Lanzarini[1,‡]

[1] *CINECA Interuniversity Consortium, Via Magnanelli 6/3, Casalecchio di Reno, Bologna, Italy*

[2] *CINECA Interuniversity Consortium, Piazza dell'Indipendenza 11/B, Roma, Italy*

Abstract

The analysis of Italian historical newspapers enhances research, accessibility, and long-term preservation by transforming digitized pages into structured and searchable resources. By fine-tuning the trained "Layout Parser" model on 3,097 manually annotated newspaper pages, we developed an Open-Source solution capable of automatically detecting the main sections of each page and generating reusable metadata. This enables scalable, automated processing of similar newspaper collections and supports more advanced forms of digital analysis.

Keywords

Newspaper Analysis, Layout parsing, Layout analysis model fine-tuning, Historical Italian newspaper, AI for Cultural Heritage, Open Source

1. Introduction

Historical newspapers are an essential part of our documentary heritage because they constitute a key to understanding the social and political context of past centuries. Although a huge number of newspapers have been published throughout history, only a small portion of them are digitally accessible. Current digitization initiatives aim to preserve them and ensure their universal access; however, the complexity of the layout structure, the variability of printing patterns, and the degradation of physical papers continue to pose significant problems in the automatic processing of historical newspapers. The digitization of historical newspapers represents, therefore, a crucial step in cultural heritage preservation, yet converting scanned images into searchable, analyzable text remains challenging.

This paper presents the work conducted by VISIT Lab at CINECA's HPC department to process a corpus of *Il Resto del Carlino* newspaper covering the 1939-1948 issues, made available by Storia e Memoria di Bologna.

While Optical Character Recognition (OCR) tools effectively handle single-column text, historical newspapers present complex multi-column layouts with varying numbers of columns, images, and titles. This requires accurate layout detection before OCR can be applied. The best candidate we identified was LayoutParser [1], but despite offering a model specifically designed for newspaper analysis, it did not produce results of the desired quality when applied to our case study.

To improve this tool's performance, we decided to enhance it through fine-tuning, which essentially means continuing the neural network's training by providing data that teaches it what we want to achieve.

IRCDL 2026: 22nd Conference on Information and Research Science Connecting to Digital and Library Science, February 19-20, 2026, Modena, Italy

*Corresponding author.

‡These authors contributed equally.

s.imboden@cineca.it (S. Imboden); a.guidazzoli@cineca.it (A. Guidazzoli); d.sforzini@cineca.it (D. Sforzini); l.mattei@cineca.it (L. Mattei); g.marconi@cineca.it (G. Marconi); f.andrucci@cineca.it (F. Andrucci); a.giannelli@cineca.it (A. Gianelli); s.caraceni@cineca.it (S. Caraceni); r.pansini@cineca.it (R. Pansini); g.fatigati@cineca.it (G. Fatigati); f.albertin@cineca.it (F. Albertin); g.baisi@cineca.it (G. Baisi); g.cardano@cineca.it (G. Cardano); m.liguori@cineca.it (M. C. Liguori); a.lanzarini@cineca.it (A. Lanzarini)

To support the production of the data needed for the fine tuning, we developed a web application (Layout Parser Data Preparation Workflow, in short LPW) specifically crafted to speed up the process and allow the contributors to work in parallel.

The production of data necessary for fine-tuning required several months of work, at the end of which a dataset of 3,097 tagged pages was created.

The fine-tuned model achieved a substantial increase in the quality of recognitions (see Table 1 - 6) and successfully identified new categories, including paragraph sections previously unrecognized.

We also present two research examples that were made possible by the availability of results from the fine-tuning work: The first, following the lead laid out by Marchand in 1985 [2], is related to the study of societal evolution through the analysis of advertising insertions: by filtering the data extracted from the newspaper *Il Resto del Carlino*, we were able to create a database of advertisements with images, text transcription, and related metadata suitable for manual or statistical analysis. The second is related to the study of language evolution: when was a new word used for the first time? When was a term first used with a specific new meaning? In this use case, we addressed a specific term, but the technique is applicable in general.

2. A Brief Overview of Online Newspaper Archives

Wikipedia provides a comprehensive catalogue of online newspaper archives, showing that most countries have published some of their newspapers online.

As stated in Wikipedia:

"Most are scanned from microfilm into pdf, gif or similar graphic formats and many of the graphic archives have been indexed into searchable text databases utilizing optical character recognition (OCR) technology. Some newspapers do not allow access to the OCR-converted text until it is proofread. Older newspapers are still in image format, but may be available as full text that can be cut and pasted and searched like born-digital newer newspapers."

We are therefore facing an evolving situation, where many sites require paid subscriptions and many have not yet completed the transformation from scans to searchable text. Among those that have, it is not obvious which technical tools were used or whether these tools are available as open-source. As CINECA, we have always focused on open-source tools to ensure a full usability by all. For this reason, we did not investigate proprietary solutions further.

Among Italian archives, we examined *Il Corriere*, *Il Sole 24 Ore*, and *La Repubblica*, which provide archive access only to subscribers, while *La Stampa* offers complete search and visualization tools with full functionality. Among regional and local archives, we observe a still-evolving situation where pages can often be located by date but only rarely by content.

3. The Case Study

The dataset used as a test-bed for the development of the workflow is made of some years of digitised *Il Resto del Carlino*.

Il Resto del Carlino, founded in 1885, is one of the longest-running Italian newspapers in circulation to date, and is generally considered the “Bolognese daily.” Specifically, the dataset examined includes copies released between 1939 and 1948. These digitalizations were kindly provided by Storia e Memoria di Bologna (History and Memory of Bologna), whose aim is to promote knowledge of the history and culture from the metropolitan area of Bologna by working together as a group of Bolognese museums and institutions.

The digitizations are in PDF format, and each file contains one-month publications of the *Il Resto del Carlino*. For the period taken into consideration,each issue numbers between 6 and 8 pages, which are usually arranged in 7 to 9 columns and distributed among 5 categories: title, text, photography,

il Resto del Carlino

UN EVENTO D'IMPORTANZA STORICA

Pio XII restituirà oggi in Quirinale la visita ai Sovrani d'Italia

Il solenne incontro richiama il più profondo interessamento della stampa internazionale

Mussolini e la Conciliazione

La risposta del Card. Maglione al messaggio del Re Imperatore

L'Ambasciatore Dino Alfieri consegna al Porporato le insegne dell'Annunziata

Figure 1: A first page of *Resto del Carlino*, 28th December 1939

drawing, and advertisements, while the contents are varied: news, press releases, weather reports, announcements, life occurrences, etc.

The body of articles tends to span multiple columns or pages, while photographs, drawings, and publications are scattered throughout the layout, as is most convenient for creating more text space and more harmony in the layout design.

However, this set of PDF files came without metadata.

The current data set is therefore an important resource, representing a challenge for those researchers eager to carry out studies through historical newspapers. Due to the multi-column layout, various types of content, and lack of structural metadata, this is a great test-bed for a system able to automatically identify and segment all elements on newspaper pages, to facilitate machine learning and digital humanities research.

4. LayoutParser

In the context of historical newspaper layout detection, the state of the art has increasingly shifted toward machine learning based approaches that can handle the structural complexity of multi-column formats, mixed text and images, and diverse typographical conventions. Traditional OCR systems that do not explicitly model layout tend to produce poor segmentation and lose semantic structure, especially for older newspapers with dense layouts and artifacts. Platforms such as Transkribus have historically provided automatic layout recognition and trainable Field Models that allow users to learn custom structural patterns for complex documents like newspapers, enabling segmentation into semantic regions such as headlines, columns, and baselines by training on annotated examples. These models have been shown to improve layout extraction by adapting to the specific characteristics of the target documents. Alongside Transkribus, modern deep learning frameworks like LayoutParser [1] represent a more general and flexible state-of-the-art approach to layout detection. LayoutParser unifies multiple pre-trained models and detection backends under a common API and employs object detection architectures based on Detectron2 [3] such as Faster R-CNN and Mask R-CNN to learn rich visual representations of document structures. These models are trained on large benchmark datasets

that include newspaper-like layouts (e.g., NewspaperNavigator) and can generalise to heterogeneous historical materials, producing robust bounding box predictions for varied layout elements. Compared to rule-based or simpler heuristic segmentation, both Transkribus' trainable field models and deep learning approaches like LayoutParser demonstrate superior accuracy and adaptability on complex newspaper pages, although deep models typically require larger annotated datasets and more computational resources.

Recent benchmarking frameworks, such as *OmniDocBench*[4], establish standardized evaluation protocols for assessing diverse PDF parsing methodologies, including newspaper-like document formats. Their findings demonstrate that pipeline-based approaches currently exhibit superior performance compared to Vision-Language Models in newspaper parsing tasks, with *MinerU* [5] achieving state-of-the-art results in this domain. However, MinerU's architecture is fundamentally optimized for PDF-centric, layout-aware document parsing, relying on structured page metadata and vector-based text/layout representations. Consequently, its performance degrades significantly when applied to plain newspaper images, which lack the requisite structural metadata and vector-encoded features inherent to PDF documents.

These approaches together reflect the leading trends in document layout analysis for historical newspapers, where learning-based segmentation is key to scalable and accurate digitisation workflows.

We adopted LayoutParser as framework for document layout analysis due to the unified and extensible environment for applying deep learning models to document images, while abstracting much of the complexity of model deployment and post processing.

LayoutParser allows for the use of models based on Detectron2, a widely adopted object detection and instance segmentation library that offers strong performance and flexibility. Leveraging Detectron2 enables the use of modern convolutional architectures and robust training and inference routines, which are particularly well suited for detecting heterogeneous layout elements in historical newspapers. From a technical perspective, the employed model follows a standard object detection architecture as implemented in Detectron2, typically composed of a convolutional backbone for feature extraction, a feature pyramid network to handle multi scale representations, and a region based detection head. The backbone network processes the input newspaper image to produce hierarchical feature maps, which are then combined by the feature pyramid network to preserve both fine grained and high level spatial information. Candidate regions of interest are generated and refined by the detection head, which performs classification and bounding box regression to identify and localize layout elements. During inference, LayoutParser orchestrates the interaction between these components and converts the raw model outputs into structured layout objects, enabling their direct use in downstream tasks such as article segmentation and text extraction within the overall pipeline. This architecture offers several advantages for historical newspaper analysis. The region based detection paradigm combined with multiscale feature representations is effective in handling complex and irregular layouts, varying font sizes, and dense page structures that are typical of historical material. The modular design of Detectron2 also allows for straightforward customization and fine tuning on domain specific data, while LayoutParser simplifies integration and experimentation. However, the approach also presents some limitations. Region based detectors are computationally demanding and may require substantial hardware resources for training and large scale inference. In addition, their performance is sensitive to the quality and consistency of annotations, which can be challenging to obtain from historical newspapers affected by degradation, skew, and printing artifacts. Despite these constraints, the chosen architecture represents a robust compromise between accuracy, flexibility, and practical usability within the proposed pipeline.

The evaluation of the quality of the recognition was performed by employing the *COCOEval* [6] module, part of the project COCO [7], through the Detectron package. The evaluation script returns the Average Precision (AP), a metric that represents the area under the precision–recall curve, summarizing how well a model ranks relevant items higher than irrelevant ones across all recall levels. The AP is calculated per class and globally.

The evaluation of the model trained on the NewspaperNavigator dataset, applied directly to *Il Resto del Carlino* collection highlights a clear issue of limited cross-domain generalization. Although

NewspaperNavigator provides large scale annotations for historical newspapers, differences in visual style, page composition, typography, and scanning quality significantly affect performance. *Il Resto del Carlino* exhibits denser layouts, different column structures, and more heterogeneous advertisement and headline styles, which are not fully represented in the source training data. This mismatch results in a noticeable drop in overall detection quality. Performance also varies widely across classes, as shown in the first column of Table 1 - 6, highlighting high scores for visually distinctive elements such as photographs, but substantially lower precision for categories like headlines and advertisements, which are more sensitive to layout conventions and font variability.

These results motivated the need for domain specific fine tuning. Fine tuning allows the model to adapt its feature representations to the distinctive visual and structural characteristics of the target newspaper, such as column density, font styles, and recurring layout patterns. By exposing the network to annotated samples from *Il Resto del Carlino*, the detection heads and higher level features can be recalibrated to better capture domain specific cues, leading to improved localization and classification performance. Without this adaptation step, even state of the art layout detection models struggle to generalize across historically and visually diverse newspaper collections.

5. Layout parser data preparation

The data required for fine-tuning Layout Parser consists of a large number of newspaper page images, each accompanied by the corresponding layout description, i.e., the decomposition of the page area (excluding margins) into rectangular areas, with each rectangle classified into one of the following categories: Title, Text, Image, Drawing, and Advertisement.

To efficiently produce these data, we developed a dedicated web application.

- no installations required
- work is automatically saved in a centralized database
- multiple operators can work simultaneously

Various measures were taken to make the work faster.

Firstly, the operator does not start from a blank page, but is presented with all the areas recognized by the previous version of LayoutParser as an initial state. Figure 2

A secondary measure involves the preliminary execution of a recognition program for page margins and lines within the page. In *Il Resto del Carlino*, lines are frequently utilized to separate text columns and to distinguish paragraphs within the same column. These lines are particularly useful during the insertion or correction of rectangles, as they function as snap points (an area that automatically attracts the mouse cursor, enabling precise positioning). The line recognizer is used during the insertion of pages into the platform, and its results can be corrected or added by an editing system purposely designed to manage cases where guideline recognition quality is suboptimal.

The capability of inserting a new rectangle with a single click proved to be particularly beneficial. From a single key input, the system detects the cursor's location and automatically calculates the new rectangle's boundaries by aligning them with the nearest guidelines or adjacent rectangles.

Supplementary editing functions allow users to remove, merge, split, or reclassify rectangles.

Annotations are exported as a JSON file with one entry per page, where each entry stores the page reference and the list of regions with their coordinates and labels.

From an architectural standpoint, the web application is containerized (Docker) and runs as cooperating services built with Svelte (frontend), PocketBase (persistence and APIs) and an NGINX reverse proxy. Svelte is a component-based web framework for building reactive user interfaces, while PocketBase provides a lightweight backend that combines persistence with customizable hooks and API endpoints. NGINX acts as a reverse proxy to route and expose web services. Additional tools like OpenSeadragon (a deep-zoom image viewer for high-resolution content) and D3.js (a JavaScript library for interactive graphics) are used on the client side for high-resolution visualization and interactive overlays. AI models are invoked via external APIs hosted on dedicated infrastructure and are not embedded in the

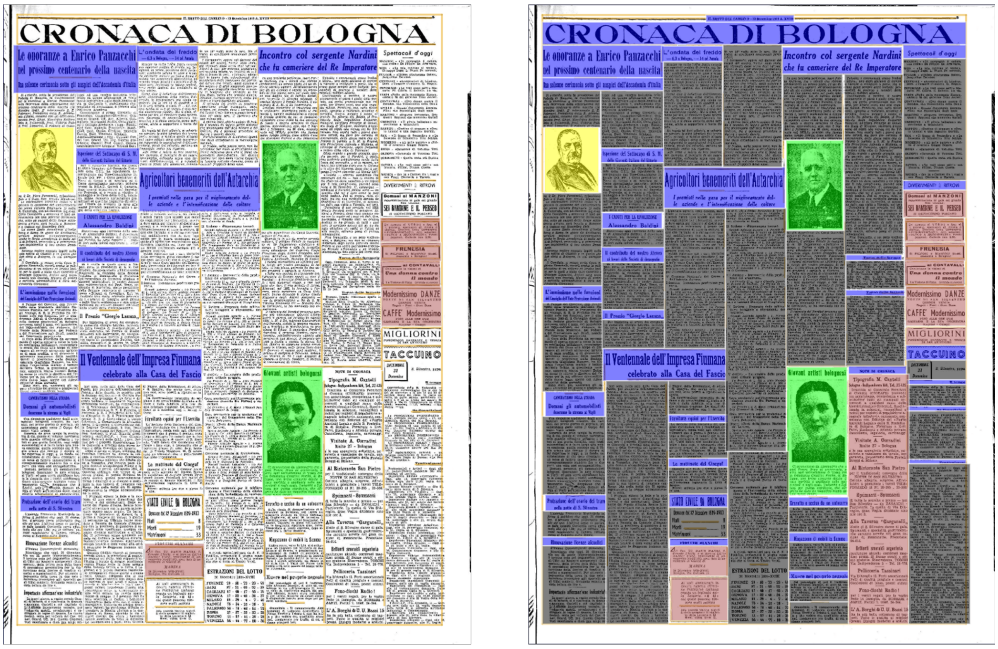

Figure 2: Screenshots from the "Data Preparation" application. On the left the page decomposition generated by the previous version of LayoutParser. On the right the correct decomposition after the operator supervision.

platform; this keeps model execution decoupled from the annotation UI while still allowing automatic proposals to be integrated into the project workflow.

5.1. A Human–AI Collaboration Environment

Dataset preparation follows an AI-assisted, human-validated paradigm: automatic algorithms generate an initial proposal, while annotators (guided by domain-defined rules) perform the final validation and correction, producing the ground truth used for fine-tuning.

Process definition. The workflow is organized as a sequence of phases that progressively transforms a scanned page into a validated training sample.

- **Automatic proposal:** before any manual work starts, the system generates a first draft through two complementary automatic steps, both executed in batch on all pages of the use case. First, a guideline-detection algorithm (heuristic-based) identifies the structural scaffold of the page -margins, column boundaries, and separators- by detecting linear alignments in the printed layout and producing editable guidelines. Second, a pre-trained layout analysis model performs region detection and classification, proposing rectangular regions labeled as Title, Article Text, Photograph, Drawing/Graphic, and Advertisement. This phase is entirely automatic and is designed to avoid "starting from a blank page": annotators do not create the layout from scratch, but instead review, correct, and complete an initial proposal, focusing their effort on boundary accuracy and label consistency rather than on exhaustive manual drawing.
- **Human correction of guidelines** (optional but decisive): when the automatic structural cues are inaccurate (e.g., misdetected column boundaries, missing separators), the annotator edits these guidelines. This step is optional because it is not always necessary; however, it becomes critical when guideline errors would cascade into poor region boundaries. The annotator can add, move, or delete guideline markers to better fit the actual page geometry.
- **Human validation and correction of regions:** annotators review every automatically proposed region and apply controlled edits: resizing to correct boundaries, deleting false positives, adding

missed regions, and resolving wrong labels. They can also split a region that incorrectly merges two units (e.g., two adjacent ads or a title+ subtitle) or merge regions that belong together (e.g., an article body fragmented into multiple blocks).
- **Human decision on usability:** pages that cannot be annotated coherently due to geometric constraints or digitization defects (e.g., strongly skewed scans) are explicitly flagged as irregular, so they can be excluded or treated separately during training.

The output of the process is therefore not "model predictions", but a **human-validated ground truth**, produced under a consistent protocol and suitable for supervised fine-tuning.

A key enabling condition is that **the workflow must be set up through a preventive domain analysis**. Before annotation starts, domain experts (e.g., historians of the press, librarians, curators, or DH researchers) define: (i) the labeling taxonomy aligned with research needs, (ii) the decision criteria for recurring borderline cases (e.g., paid notices vs editorial text), and (iii) the granularity of units that should be learnable by the model. This step is crucial: without shared domain decisions, different annotators would apply different interpretations, producing label noise that directly harms fine-tuning and reduces generalization.

Why "scrupulous" tagging rules matter. In supervised layout analysis, annotation consistency is as important as annotation quantity. In our setting, the model is fine-tuned to learn both **semantic categories** (what a region is) and **geometric conventions** (where its boundaries should be). If annotators are inconsistent (e.g., sometimes including decorative rules inside boxes, sometimes excluding them; sometimes absorbing a photo headline into the image, sometimes separating it) the model receives contradictory signals and either fails to converge or learns unstable heuristics that do not transfer across pages. Scrupulous guidelines therefore serve three purposes: (1) they make human corrections faster by reducing uncertainty ("what should I do in this case?"), (2) they reduce inter-annotator variance, and (3) they encode domain knowledge into the dataset in a form the model can learn. In practice, the AI contributes speed and coverage, **but the dataset becomes training-grade only when annotators enforce the rules systematically.**

Most relevant tagging rules (with examples). The following principles were treated as high-priority because they recur frequently and have a strong effect on downstream reuse (article segmentation, advertisement extraction, and visual object retrieval):

- **Titles: separate logical structure, not mere line breaks.** Titles should be segmented into distinct regions only when there is a real typographic hierarchy (e.g., a headline clearly separated from a subtitle). A simple wrap onto a second line does not automatically imply two regions. Paragraph headings inside an article are still Title regions. In advertisements, typographic headings (brand names, slogans) are treated as titles only if they function as a headline and are visually distinct; purely decorative typography should not be promoted to a title region.
- **Exclude decorative ornaments whenever they do not carry content**. Separators and ornamental lines frequently appear between articles and inside headers. Annotators should avoid including these elements in content regions (Title/Text/Ad) unless they are integral to the semantic unit (e.g., a boxed table-like structure that defines the ad boundary). This prevents the model from learning "ornament = content".
- **Photographs include captions, but not external headlines.** The Photograph region should include its caption when the caption is visually attached (common in newspapers). However, when a photograph has a headline above it that acts as an article title, that headline should be labeled as Title, not absorbed into the image region. This separation is relevant for later retrieval: it preserves the distinction between visual content, descriptive caption, and editorial heading.
- **Drawings/graphics include captions and informational markers.** Maps, charts, plots, and illustrations are labeled as Drawing/Graphic. If a caption is present and clearly belongs to the figure, it is included in the same region (so the figure remains a coherent unit). Small typographic markers may be included only when they play an informational role (e.g., map symbols, scale indicators), but purely decorative flourishes should be excluded.

- **Advertisements: separate items, keep the ad atomic.** When multiple ads appear in a grid, column, or classifieds-like arrangement, each ad should be separated into its own region. This supports use cases where ads are extracted as discrete objects with metadata. Inside a single advertisement, however, annotators do not further subdivide drawings vs text: the ad remains an atomic unit, preserving the integrity of the paid message and avoiding inconsistent micro-segmentation.
- **Paid notice heuristic for ambiguous "text vs advertisement".** Some blocks resemble editorial notices (e.g., family announcements, subscription tariffs, public communications). To ensure consistent labeling, the adopted rule is operational: if the notice is paid to be published, it is labeled as Advertisement; otherwise it is Article Text. This domain-driven heuristic is simple enough to apply consistently and matches socio-economic interpretations of newspaper content.
- **Dates and marginal metadata: label the content, not the frame.** Dates or small publication markers are labeled as Article Text if they are meaningful textual content, but annotators avoid including surrounding framing boxes when the frame is merely graphic scaffolding.
- **Rectangular constraints and "irregular page" handling.** Because layout regions are axis-aligned rectangles, strongly skewed pages can make faithful segmentation impossible (e.g., trapezoidal columns). In these cases, annotators do not force a poor approximation by splitting columns into many small rectangles: an approach that would confuse the model. Instead, the page is flagged as irregular so it can be excluded or handled separately.

Taken together, these phases and rules operate a clear division of labor: **the algorithm proposes, humans decide**. The AI accelerates the creation of candidate structures; domain experts define what "correct" means for the research goals; and trained annotators, guided by scrupulous rules, deliver the consistent ground truth that makes fine-tuning effective and reusable across similar newspaper collections.

6. Fine-tuning

To address the observed generalization gap, we performed a domain specific fine tuning of the layout detection model using computational resources provided by the Leonardo supercomputer, exploiting a single node of the Booster partition. To fine-tune the model we used 3,097 manually annotated newspaper pages, with a train-test partition of 80-20. Model adaptation was performed using the official Detectron2 software development kit, ensuring full compatibility with the underlying architecture and training procedures. To further analyze the impact of pre training data and label configuration on layout detection performance, we conducted three distinct fine tuning experiments using a Faster R CNN architecture within the Detectron2 framework. All experiments shared the same model configuration and training protocol, differing only in the source of the initial weights and in the inclusion of text regions among the target classes. In the first setup, the model was initialized with weights pre trained on the NewspaperNavigator dataset and fine tuned without including text regions as explicit detection targets. This configuration focuses exclusively on higher level layout elements such as photographs, advertisements, and headlines. In the second setup, we returned to the NewspaperNavigator initialization but extended the label set to include text regions, enabling the model to jointly learn both structural and textual layout components. Finally, the third setup combined ImageNet pre training with the extended label configuration including text regions. These three configurations enable a systematic comparison between document specific and generic pre training, as well as an assessment of how explicitly modelling text regions influences the detection of other layout elements.

Based on the comparative evaluation of the three fine tuning configurations, we ultimately opted for the model initialized with ImageNet pre trained weights and trained with the extended label set including text regions.

Metric %	Before FineTuning	FT 1	FT 2	FT 3
AP	24.59	71.02	71.84	71.86
AP-Photograph	55.74	86.01	84.12	82.41
AP-Comics/Cartoon	8.75	57.72	56.02	59.11
AP-Headline	13.78	68.30	66.64	67.06
AP-Advertisement	20.10	72.02	66.97	65.72
AP-Text	–	–	85.43	85.01

Table 1
Layout detection performance of: original NewspaperNavigator model, fine tuned NewspaperNavigator-notext model (FT 1); fine tuned NewspaperNavigator-text model (FT 2); fine tuned ImageNet-text model (FT 3)

7. Use Cases: "Printed ads" and "Piovra"

In order to guide the development of the solution, a series of User Journeys were gathered among researchers to check which uses of a workflow involving the LayoutParser applied to historical newspapers might be desirable the most. Among them, two have been tested up to now.

The first one is about the analysis of printed advertisements; the second one is about a single word and the contexts of its use.

7.1. Research and analysis of printed advertisements

One of the User Journeys was about "comparative analysis of historical trends through the socio-economic evaluation of printed advertisements in daily newspapers". In this case, the desired output aimed at obtaining support for research and analysis of the ads and the possibility of downloading them along with their related metadata. The requirements were:

- display of automatically selected items of interest (printed advertisements – images and text);
- interaction with the selected items to adjust any incorrect selection;
- download of images (RGB or black and white) and text associated with the images;
- consultation/download of metadata associated with objects of interest in formats that are easy to use for subsequent analysis.

Newspaper advertisements should be searchable based on a series of information: date of publication; page of publication, dimensions (absolute and normalised with respect to the number of columns in the newspaper); position on the page [8] (based on a division into 4 quadrants); image description; brand of the advertised product or service (if applicable); transcribed text. All this information, and downloaded images, should be organised in a way that is easy to consult and modify.

The User Journey was inspired by historical research previously conducted using traditional methods [9]. In this way it could better help in order to define truly useful objectives and to evaluate the actual benefits obtainable through the use of Layout Parser and a workflow integrated with other artificial intelligence models.

Once the fine-tuning phase on LayoutParser was completed, an initial test was carried out on a one-month selection of *Il Resto del Carlino*. Based on the bounding boxes generated by LayoutParser, subsequent segmentation was then applied using an algorithm. Finally, a multimodal LLM, specifically QWEN3, was applied to the individual images of each advertisement, with a prompt aimed at generating a transcription of the text in the advertisement, a description of it, the identification of the brand promoting goods and services and their belonging to a specific product area. The product areas were defined using the list with descriptions presented in [9].

While the transcriptions were generally correct, as were the descriptions, brand identification and product categorisation fell well below 50%, with errors mainly concentrated in job advertisements. However, while being aware that technical biases can skew historical analyses [10], these errors do not detract from the excellent result, which offers researchers an excellent level of segmentation and

extraction of individual advertisements, with date matching and size definition (fig. 3). With the provision of an interface to manage the results, it is still possible for researchers to intervene in the correction of the results, further increasing the amount of correct information obtainable even from a rather large dataset.

Figure 3: Screenshot taken from the database of advertisements with images, text transcription, and related metadata suitable for manual or statistical analysis (nocodb)

7.2. "Piovra", the use of a word and its context

One of the main challenges encountered by historical researchers when working with digital data is the difficulty of generating new research questions—that is, shifting from a traditional epistemological framework to one informed by the full potential of digital methods. This shift concerns not only the scale of analysis but also the very nature of historical inquiry, enabling the systematic study of cultural phenomena that were previously accessible only through limited or exemplary cases.

The research proposed by professor Roberto Balzani addresses this challenge through the original case of the "piovra" (octopus), a term absent from the Italian lexicon until Victor Hugo made it famous by associating it with malevolent traits. From that moment, the word and its powerful figurative representation—often caricatural and satirical—spread across literary, political, social, and economic discourses. Newspapers offer a privileged source for tracing this diffusion in Italian common use between 1866 and 1900, as they played a central role in popularizing both the term and its imagery through articles and serialized fiction.

The study focuses on newspapers of different political orientations and geographical origins, analyzing the spatial placement of the term within page layouts, its semantic associations with selected indicators, and its appearance in prominent areas of the front page as evidence of symbolic consolidation. Methodologically, the research requires querying for an optimal dataset of historical newspapers in a significative temporal range and applying an ad hoc workflow in order to automatically identify the occurrences.

The ultimate goal of this User Journey is to construct a semantically enriched dataset in which the structured articulation of meanings derived from the lemma itself constitutes an interpretative outcome.

To support this objective, a dedicated processing pipeline was designed to integrate image-based analysis, automated text recognition, and semantic indexing within a single workflow. The pipeline operates on large collections of digitized newspaper pages, systematically segmenting them into meaningful regions and linking lexical occurrences to their precise spatial contexts. This approach ensures that each detected instance of the lemma is embedded within a rich set of interpretative metadata, including page topology and visual prominence. The following section details the architecture of this pipeline, its execution logic, and the computational resources that enable large-scale, reproducible analysis of historical press materials

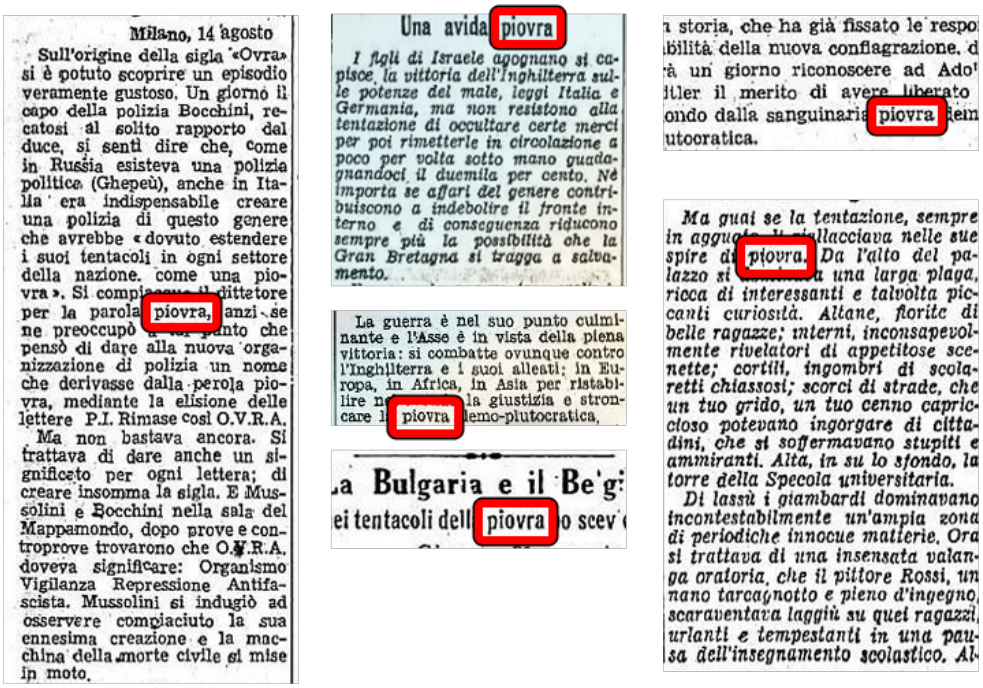

Figure 4: Some examples of lemma "piovra" found in the dataset.

From a technical standpoint, the pipeline is structured as a multi-stage process centered on a fine-tuned layout analysis model specifically trained on historical newspapers. The first stage operates directly on the digitized page images and applies to the finetuned LayoutParser model.

The model outputs a structured JSON representation of the page, in which each detected region is classified according to its functional role (e.g., title, text, advertisement, photograph, illustration). This representation constitutes the backbone for all subsequent processing steps.

In the second stage, only regions classified as textual content are selected for optical character recognition. Each text region is processed independently using a Keras-based OCR model, chosen specifically for its ability to return word-level bounding boxes alongside the recognized textual strings. This design choice enables a precise alignment between lexical data and spatial information, preserving the positional context of each word within the page layout. As a result, every occurrence of the lemma "piovra" can be localized not only at the page level but also within specific textual blocks and visual hierarchies.

Finally, the recognized words, together with their bounding boxes, region identifiers, and page-level metadata, are ingested into a database 4. This step transforms the newspaper collection into a fully searchable corpus. Through this pipeline, the fine-tuned layout analysis model acts as an enabling layer that bridges image-based document structure and semantic exploration, allowing historical interpretation to emerge from the systematic correlation between language, space, and visual organization.

8. Conclusions and Future Perspectives

This work demonstrates how AI-powered layout detection, combined with human expertise and high-performance computing resources, can unlock the full potential of digitized historical newspapers for research, enabling content-based search, statistical analysis, and specific historical investigations such as advertising trends analysis and semantic evolution studies.

In our case, fine-tuning activities led to a significant improvement in the model's performance: we went from AP 24.59% to AP 71.86%, which correspond approximately to the correct recognition of the 95% of the page area. More importantly, we were also able to extend the categories of recognised areas to include the "paragraph" area, which was not previously included.

At the moment, we have performed LayoutParsing on all the newspapers made available to us and OCR has been performed on individual areas of each page, effectively digitising all text. The determination of the hierarchical structure of the page, i.e. the link between the various areas that make up an article and the corresponding title, is still missing.

While awaiting completion of the digitisation process as described above, we have received further requests for support from researchers wishing to carry out specific research based on this data. The use cases implemented to date already clearly demonstrate how, for example, it is possible to facilitate and speed up work on advertisements, so that statistical analyses can be carried out and their evolution over the years studied; and how to facilitate studies on the evolution of the Italian language, contributing, for example, to identify the moment when a specific new term appears, or when a new meaning is associated with an existing term. The areas of research that will benefit from these tools are undoubtedly very broad.

As future work, we plan to explore alternative detection architectures to further improve robustness and accuracy on historical newspaper layout detection. In particular, we intend to evaluate single stage detectors, as YOLO[11] ,and more recent transformer based models for document layout analysis, which may offer advantages in terms of global context modelling and scalability.

These directions will allow us to assess the trade off between computational cost, generalization capability, and detection performance, and to identify architectures that are better suited for the variability and complexity of large scale historical newspaper collections.

Acknowledgments

This work has been made possible thanks to the funding provided by the Italian Ministry of Culture (MiC) – Digital Library, which operates under the Istituto Centrale (IC) for the digitization of cultural heritage, itself part of the Directorate General for Digitization and Communication, within the framework of the Italian National Recovery and Resilience Plan (PNRR), under Mission 1: “Digitalisation, innovation, competitiveness, culture and tourism”, financed by the European Union – Next GenerationEU. The authors would like to express gratitude to Prof. Roberto Balzani, Professor at the University of Bologna (Alma Mater Studiorum), for the academic guidance that encouraged the development of this study to include not only layout parsing but also the use of OCR analysis. Finally, we would like to express our gratitude to Storia e Memoria di Bologna for providing us with the opportunity to utilize the digitalizations of the *Resto del Carlino* and for their coordination efforts in preserving and promoting local historical heritage.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] Z. Shen, R. Zhang, M. Dell, B. C. G. Lee, J. Carlson, W. Li, Layoutparser: A unified toolkit for deep learning based document image analysis, arXiv preprint arXiv:2103.15348 (2021).

[2] R. Marchand, Advertising the American Dream: Making Way for Modernity, 1920–1940, Berkeley: University of California Press, 1985.

[3] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, R. Girshick, Detectron2, https://github.com/facebookresearch/detectron2, 2019.

[4] L. Ouyang, Y. Qu, H. Zhou, J. Zhu, R. Zhang, Q. Lin, B. Wang, Z. Zhao, M. Jiang, X. Zhao, J. Shi, F. Wu, P. Chu, M. Liu, Z. Li, C. Xu, B. Zhang, B. Shi, Z. Tu, C. He, Omnidocbench: Benchmarking diverse pdf document parsing with comprehensive annotations, 2025. URL: https://arxiv.org/abs/2412.07626. arXiv:2412.07626.

[5] B. Wang, C. Xu, X. Zhao, L. Ouyang, F. Wu, Z. Zhao, R. Xu, K. Liu, Y. Qu, F. Shang, B. Zhang, L. Wei, Z. Sui, W. Li, B. Shi, Y. Qiao, D. Lin, C. He, Mineru: An open-source solution for precise document content extraction, 2024. URL: https://arxiv.org/abs/2409.18839. arXiv:2409.18839.

[6] cocodataset.org, Coco - detection evaluation, https://cocodataset.org/#detection-eval, 2021.

[7] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, P. Dollár, Microsoft coco: Common objects in context, 2015. URL: https://arxiv.org/abs/1405.0312. arXiv:1405.0312.

[8] M. Wevers, Mining Historical Advertisements in Digitised Newspapers, De Gruyter Oldenbourg, Berlin, Boston, 2023, pp. 227–252. URL: https://doi.org/10.1515/9783110729214-011. doi:doi:10.1515/9783110729214-011.

[9] M. C. Liguori, North and south: Advertising prosperity in the italian economic boom years., Advertising & Society Review 15 (2015).

[10] K. Venglarova, R. Adam, W. Mölzer, S. Balasubramanian, J. Kleinert, M. Füllsack, G. Vogeler, Who advertises in newspapers? data criticism in mining historical job ads, in: CHR 2024: Computational Humanities Research Conference, 2024.

[11] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, 2016. URL: https://arxiv.org/abs/1506.02640. arXiv:1506.02640.