PIVAJ: an article-centered platform for digitized newspapers

Pierrick Tranouez, Stéphane Nicolas, Julien Lerouge, Thierry Paquet.
LITIS – University of Rouen
FRANCE
PIVAJ@litislab.eu

Abstract

PIVAJ is a platform for archived digitized newspaper emphasizing articles: extracting them from digitized documents by automated page layout analysis, OCRing them, indexing their text transcription to allow users to search for content. Crowdsourcing is used to improve the quality of the indexing, by correcting the transcription and by tagging articles with keywords. The platform has been used to give Web access to 550 000 articles generated from a digitized local newspaper. Current developments include further improvements to its OCR as well as graphical interfaces for the management of the platform.

Articles as elementary parts of information in newspapers

Newspapers are periodical publications containing news and feature articles, generally printed on low quality paper. The link among the information it contains is the periodicity itself: news are relevant to the day or the week of the issue, feature articles may be relevant to a wider timespan but are still generally closely linked to the period of printing. The common thread is temporal, and is consumed thus by the readers buying the newspaper.

This immediate view of the present comes later on to be of use for other people, interested in these events of the past: historians, both pro and hobbyists, genealogists, or just individuals curious of local history. But for them, the dates of printing, the issue itself as a container of articles, often lack relevance, except as a way of context. They want to retrieve information; there is a theme to their research: they may need to parse 20 articles spread in 20 issues over 10 years of publication to get the data they want.

PIVAJ is a platform for archived digitized newspaper built to help people easily access the precise information they search, even though it may be disseminated in many pieces. Articles are the unit of this information. Sections and articles are therefore the core materials PIVAJ tools manage and produce: extraction, analysis, transcription, indexing and visualization.

Newspapers layout

Historical newspapers have complex structures that may vary across the different layouts encountered. These structures provide much context to help the processing of the information contained by the newspapers, this is the reason why we aim at correctly extract and represent them in PIVAJ.

In general, the headings (of different levels) and the separators, which are detected during the automatic layout extraction step, are the most helpful clues for extracting the hierarchical structure of a newspaper.

Newspapers may be composed of:

- A banner, which includes the name of the newspaper, and two ears (only on the title page)
- Headers, footers and margins
- Multiple sections (often delimited by horizontal separators)

Newspaper *sections* are the containers of the articles. They may span on multiple pages of a single issue, and may have a heading, especially when they recurrently appear over the issues. They are often divided into several columns by vertical separators. See Figure 1 for a sample composition of a title page.

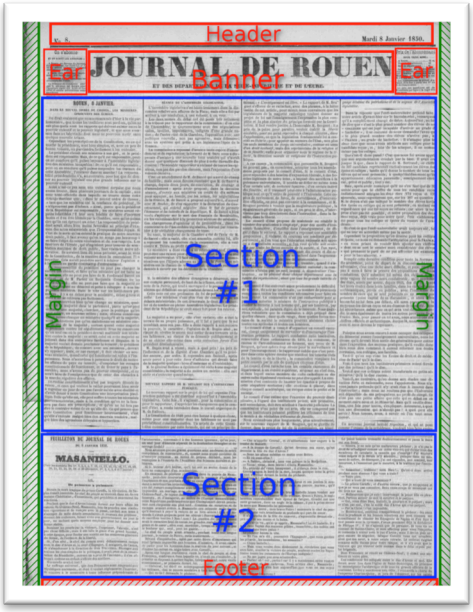


Figure 1: Composition of a title page. Articles are in sections.

Besides, articles also have complex and varying structures. As we defined them in PIVAJ, articles may contain:

- A heading, which may span on multiple lines
- A subheading, which is a heading of a lower level
- Parts, which are marked either by horizontal separators or by consecutive subheadings
- Paragraphs containing body text
- Illustrations, tables, captions…
- Annotations (signature, dates, footnotes)

An article part may contain a heading, (sub)parts, paragraphs and other contents, just like an article does. Thanks to its recursive

nature, our definition of an article is generic enough to be applied on a variety of historical newspapers. However, in general, subparts (or deeper structures) are not needed. See Figure 2 for a sample composition of an article.


Un Assaut Franco-Italien

Kirchoffer et Lancia di Brollo. — Une belle Séance d'armes. — La Victoire du Maître français.

LA RENCONTRE

Kirchoffer

Lancia di Brollo

Figure 2: Article example, with different parts, subheading, illustrations etc.

Automatic layout extraction

Once digitized, a newspaper is an organized collection of pictures. Each of these pictures from the point of view of the computer is just a collection of colored dots laying together rows after rows. This is far from the computational usability of native digital documents, such as a Word document or a Web page, which contain digital text and other textual metadata that allow indexing, searching, using multiple layouts etc. Several steps must be taken to grant those digitized documents some properties of these native digital documents.

First, PIVAJ must understand the layout of each page of a newspaper. To this end, it uses several statistical machine learning algorithms. This learning characteristic implies the necessity to provide a certain amount of representative pages with their detailed *ground truth*, which are images where all the relevant information (i.e. all the texts, headings with their appropriate level, the separators, the pictures, the captions etc.) have been delimited as blocks by hand. PIVAJ is then able to learn from these examples how to automatically label the different parts of the images as textual content (body text, headings of different levels, captions etc.) or graphical content (separators, pictures etc.).

Pictures are detected separately using an algorithm based on Random Forests [1], working on the grey levels histograms. Other content (text and separators) are detected by an algorithm based on Conditional Random Fields, as described in depth in [2]. These two analyses are then fused.

The steps of the automatic page layout extraction are as follows:

1. First PIVAJ labels textual and graphical blocks, based on what it learnt from the ground truth examples.
2. From this step, it then builds a grid, assembling the visible separators and creating blank ones where needed (e.g.: between columns of text).
3. Using this grid it infers a first reading order, which allows it to use a second machine learned discrimination inside text zones to recognize the headings with their appropriate levels.

Those parts are then assembled in sections and articles. However, this assemblage may vary between different newspapers, this is why we made it adaptive using regular expressions that describe the way heading, body text, separator, and graphical blocks detected in the previous stages should be aggregated to form a section or an article. The full process (labeling and assembling) is automatic. Figure 3 gives an example of this process, from the initial page to the illustrative colored representation of the detected articles, including the underlying hierarchy of sections. Figure 4 details this process at the article level.

A graphical software tool is included in PIVAJ so that an operator may build the ground truth, manage the learning process, and then launch the analysis, without requiring any knowledge about the underlying image analysis and machine learning processes.

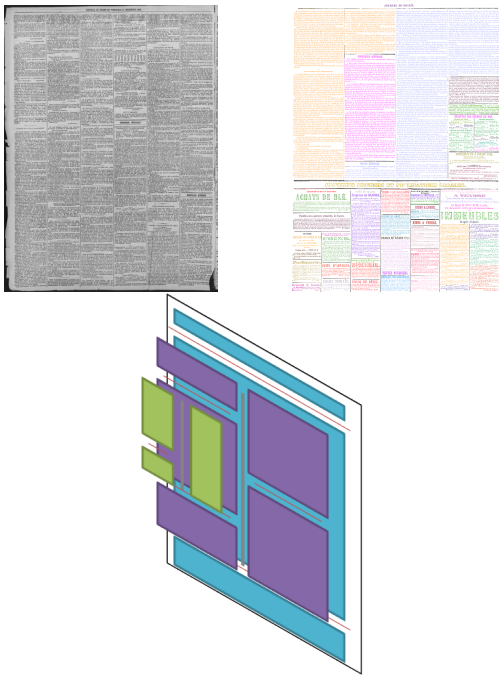
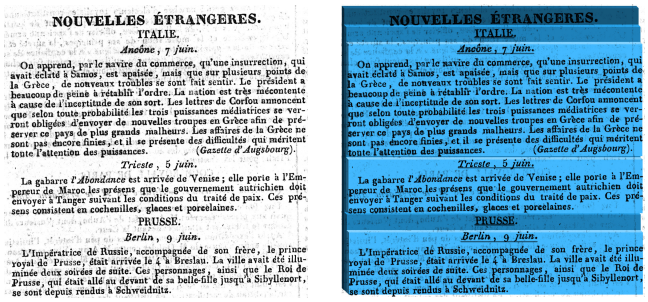


Figure 3: extracting the parts of the layout from the digitized image by automatic analysis


NOUVELLES ÉTRANGERES.

ITALIE.

Ancône, 7 juin.

On apprend, par le navire du commerce, qu'une insurrection, qui avoit éclaté à Samos, est apaisée, mais que sur plusieurs points de la Grèce, de nouveaux troubles se sont fait sentir. Le président a beaucoup de peine à rétablir l'ordre. La nation est très mécontente à cause de l'incertitude de son sort. Les lettres de Corfou annoncent que selon toute probabilité les trois puissances médiatrices se verront obligées d'envoyer de nouvelles troupes en Grèce afin de préserver ce pays de plus grands malheurs. Les affaires de la Grèce ne sont pas encore finies, et il se présente des difficultés qui méritent toute l'attention des puissances. (*Gazette d'Augsbourg*).

Trieste, 5 juin.

La gabarre *l'Abondance* est arrivée de Venise; elle porte à l'Empereur de Maroc les présens que le gouvernement autrichien doit envoyer à Tanger suivant les conditions du traité de paix. Ces présens consistent en cochenilles, glaces et porcelaines.

PRUSSE.

Berlin, 9 juin.

L'Impératrice de Russie, accompagnée de son frère, le prince royal de Prusse, était arrivée le 4 à Breslau. La ville avait été illuminée deux soirées de suite. Ces personnages, ainsi que le Roi de Prusse, qui était allé au devant de sa belle-fille jusqu'à Sibyllenort, se sont depuis rendus à Schweidnitz.

NOUVELLES ÉTRANGERES.

ITALIE.

Ancône, 7 juin.

On apprend, par le navire du commerce, qu'une insurrection, qui avoit éclaté à Samos, est apaisée, mais que sur plusieurs points de la Grèce, de nouveaux troubles se sont fait sentir. Le président a beaucoup de peine à rétablir l'ordre. La nation est très mécontente à cause de l'incertitude de son sort. Les lettres de Corfou annoncent que selon toute probabilité les trois puissances médiatrices se verront obligées d'envoyer de nouvelles troupes en Grèce afin de préserver ce pays de plus grands malheurs. Les affaires de la Grèce ne sont pas encore finies, et il se présente des difficultés qui méritent toute l'attention des puissances. (*Gazette d'Augsbourg*).

Trieste, 5 juin.

La gabarre *l'Abondance* est arrivée de Venise; elle porte à l'Empereur de Maroc les présens que le gouvernement autrichien doit envoyer à Tanger suivant les conditions du traité de paix. Ces présens consistent en cochenilles, glaces et porcelaines.

PRUSSE.

Berlin, 9 juin.

L'Impératrice de Russie, accompagnée de son frère, le prince royal de Prusse, était arrivée le 4 à Breslau. La ville avait été illuminée deux soirées de suite. Ces personnages, ainsi que le Roi de Prusse, qui était allé au devant de sa belle-fille jusqu'à Sibyllenort, se sont depuis rendus à Schweidnitz.

Figure 4: Zoom in at the article level

Automatic transcription

Once the layout understood by the software, an OCR is applied to the detected text zones for transcription. Although an OCR is being included in PIVAJ, any other OCR can be substituted and easily plugged in the process. For example, we have done extensive testing with Google Tesseract and OCRopus. An ALTO description of the page is produced at this step. Additional information relative to the hierarchy of the sections, the articles and the reading order are input into the database for later use by the indexing and visualization components. This information is also input into the METS description of each newspaper issue.

Retrieving information: Web visualization, indexing, crowdsourced corrections

Reading archived newspapers on the Web

The information contained in the digitized newspapers can be accessed through the Web in different ways.

The most basic one is image visualization accessed by date or issue number. Each issue of a newspaper can be accessed and read on screen. We use IIpImage to manage images in the TIFF pyramidal format, which allows for both very fast and very detailed rendering of an image, as illustrated in Figure 5. Only the part of the images necessary at the scale of the viewing is thus downloaded.

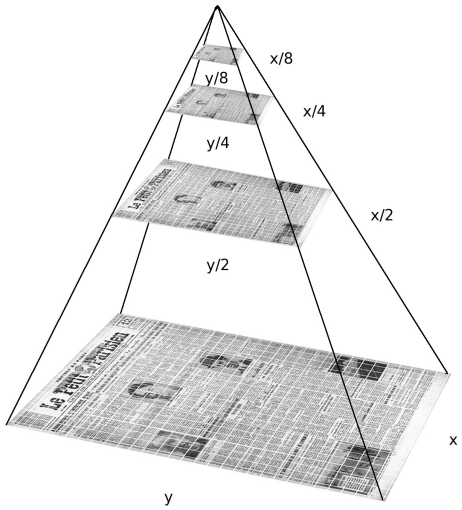


Figure 5: TIFF pyramidal format

But information can also be accessed through text queries. Metadata and the complete transcriptions are indexed (using Apache Lucene) so that relevant articles are retrieved for a given query. Different criteria can then be used to refine the initial answer (newspapers, categories, decades, years, months), as illustrated in Figure 6. Furthermore, another text search can be queried among these first answers.

Figure 6: searching information

As pictures have been labeled by the automatic layout extraction, we index them according to their caption and the text of the article they are included in. Users can therefore separately ask for pictures (graphics, photos etc.) using a textual query.

Improving the transcription through crowdsourcing

The quality of the search is very dependent on the quality of the text transcription. Archived newspapers can be centuries old, thus giving most OCR a hard time. The digital text associated to the image can in these cases be close to unusable.

We built an interface to let users correct a defective transcription. Its look and feel is similar to NLA's Trove [3], as both evolved from NLA's once open source newspaper beta service. Underneath everything is different: database, software architecture, indexing and visualization tools.

Figure 7: Reading an article in PIVAJ

The image of the corrected article and the text are zoomed in and aligned to allow for an easy correction. In our experiments,

users have concentrated on the correction of named entities, one of the particular weaknesses of most OCR, and thus improved considerably the relevance of search requests, especially since nearly half the research led on newspaper archive are motivated by genealogical problem, according to a survey conducted amongst our users.

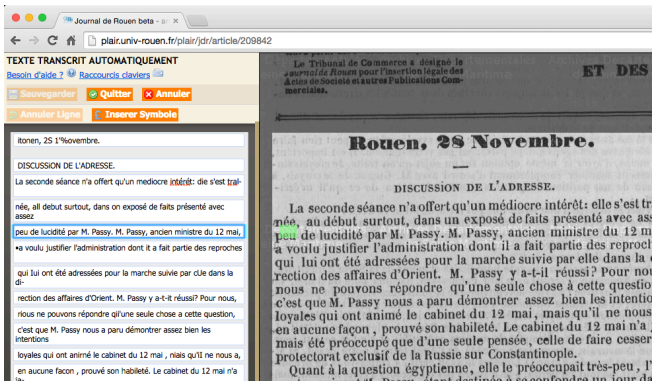


Figure 8: Transcription correction

Users can also tag keywords to the articles. These tags can be private or public. They can afterward be used as search criteria.

Prototype and experiments

Since [4], many experiments were carried out which led to several improvements.

PIVAJ is a project developed by the University of Rouen. It led in 2013 to a collaboration with the archives of Seine-Maritime, to give access to 50 years of a local daily newspaper called Le Journal de Rouen. This represents 100 000 pages and about 550 000 articles. It is online at http://plair.univ-rouen.fr/. An average of 1500 users connect every month, and about 40 000 lines have been corrected by crowdsourcing since the beginning. 70% of these users arrive by direct Google query: they were not looking for the Journal de Rouen, they were searching for information and found it in the articles of the newspaper PIVAJ gave them access to.

In 2013 our crowdsourcing correction tool has been tested in collaboration with the National Library of the Netherlands during the OCR pilot carried by European project IMPACT. Our platform, referred to under the name of its mother project PlaIR, was ranked first out of 5 different tools that were evaluated during this campaign [5].

Current development resumed in September 2014 and includes improving the efficiency of the layout analysis beyond what we did in [1]. We are also building extensive graphical user interfaces for all the elements of the software, to allow for an easy deployment and management of the platform.

We were provided access by the Bibliothèque Nationale de France (BnF) to a host of different newspapers from the XIXth and early XXth century. Preliminary analysis of our results by the BnF on 7 issues of Le Journal des Débats, from 1815 to 1850 concluded on a 93% quality in their metrics (number of fully extracted articles plus number of partially extracted articles times 0.5, the whole divided by the total number of articles to be extracted, as estimated by a BnF editor).

Figure 9: Layout extraction from Le Journal des Débats

References

[1] L.Breiman. "Random forests." *Machine learning* 45.1 (2001): 5-32.

[2] T. Palfray, D. Hebert, S. Nicolas, P. Tranouez, and T. Paquet. *Logical segmentation for article extraction in digitized old newspapers*. In Proceedings of the 2012 ACM Symposium on Document Engineering, DocEng '12, pages 129–132, New York, NY, USA, 2012. ACM.

[3] R. HOLLEY. *Trove: Innovation in access to information in Australia*. In Ariadne, Issue 64, 2010.

[4] T. Palfray, S. Nicolas, T. Paquet and P. Tranouez. *"PlaIR": A System to Provide Full Access to Digitized Newspaper Archives* in Archiving 2012.

[5] http://www.impact-project.eu/uploads/media/IMPACT_D-EXT2_Pilot_report_KB.pdf

Authors Biography

PIVAJ is a spinoff of a broader information indexing and retrieval project called PlaIR. It was developed from 2010 to 2012 under PlaIR's funding (regional and European) and benefits from a specific regional funding since 2014.

Dr. Pierrick Tranouez heads this second phase of PIVAJ with Julien Lerouge as the lead developer of the platform. Dr. Stéphane Nicolas is a specialist of document layout analysis. Prof. Thierry Paquet co-headed the PlaIR project and is the current director of the LITIS laboratory, a computer science research unit spanned across the University of Rouen, University of Le Havre and INSA of Rouen.