https://doi.org/10.2352/issn.2168-3204.2023.20.1.10

# Pivan: a Web platform for document annotation

*Thomas CONSTUM, Florian BEBIN, Pierrick TRANOUEZ, Thierry PAQUET; LITIS – University of Rouen Normandie; France*

## Abstract

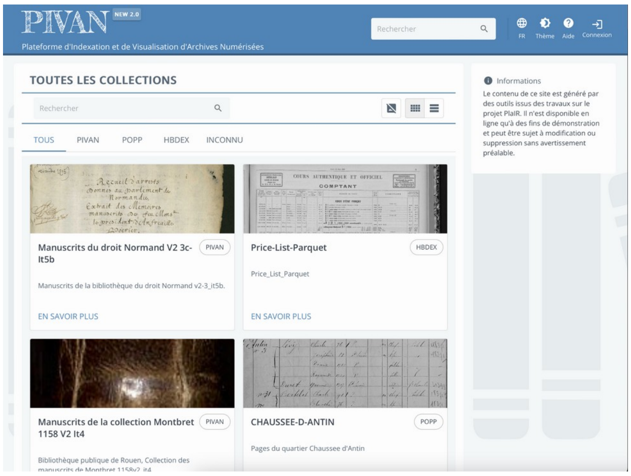
*The Pivan web platform is an open-source tool for managing different stages of automatic document processing, such as layout analysis, transcription, and named entity recognition. It allows for the visualization of document segmentation, transcription at the line or paragraph level, and annotation of named entities. Pivan's web-based nature makes it perfectly suited for collaborative annotation and offers a smooth experience, even for small machines or connections. It is based on up-to-date web technologies, it includes a comprehensive API, and it can be easily deployed via Docker.*

## Introduction

In the field of automatic document processing, much attention is paid to improving layout analysis, handwriting recognition, and information extraction methods. Nowadays the best methods rely on Supervised Machine Learning, mainly Deep Learning. In these methods a system is trained to perform a classification or regression task by being presented numerous examples of what a correct classification should be. This supposes the existence of what is called ground-truth, which is the result of a human performance of the task the system will be trained to perform. In practice, this means a significant part of the projects is devoted to the production of such ground truth. It is therefore crucial to use an easy-to-use and efficient tool to reduce the time needed for this task and to keep annotation errors to a minimum. The same tool can be used to visualize the result of the work of the system once it has been trained, to better understand its performance, or simply to correct what it did wrong.

## Use cases

Pivan is an open-source web platform aimed at managing different stages of document collections that are processed automatically by a machine. It is a spinoff of the online part of PIVAJ [1], a platform for archived digitized newspapers emphasizing articles segmentation, recognition, and indexing. The main use cases of Pivan are manual production or correction of layout analysis, transcription and named entities recognition. Its input/output format is ALTO/METS files. Figure 1 shows some examples of collections supported by Pivan.



Informations

Le contenu de ce site est généré par des outils issus des travaux sur le projet PlaIR. Il n'est disponible en ligne qu'à des fins de démonstration et peut être sujet à modification ou suppression sans avertissement préalable.

Manuscrits du droit Normand V2 3c-It5b

Manuscrits de la bibliothèque du droit Normand v2-3_it5b.

Price-List-Parquet

Price_List_Parquet

Manuscrits de la collection Montbret 1158 V2 It4

Bibliothèque publique de Rouen, Collection des manuscrits de Montbret 1158v2_it4

CHAUSSEE-D-ANTIN

Pages du quartier Chaussee d'Antin

***Figure 1***. *An overview of the diversity of collections supported by Pivan*

### Layout analysis

PIVAN allows for the visualization of image layout segmentation and labeling of document images. The segmentation labels are represented as bounding boxes for the different elements of the document structure: page, paragraph, and line. (See Figure 2) They can be structured according to a tree structure defined in the XML METS files imported into Pivan. The METS/ALTO format allows Pivan to support various types of structures such as tabular structures.

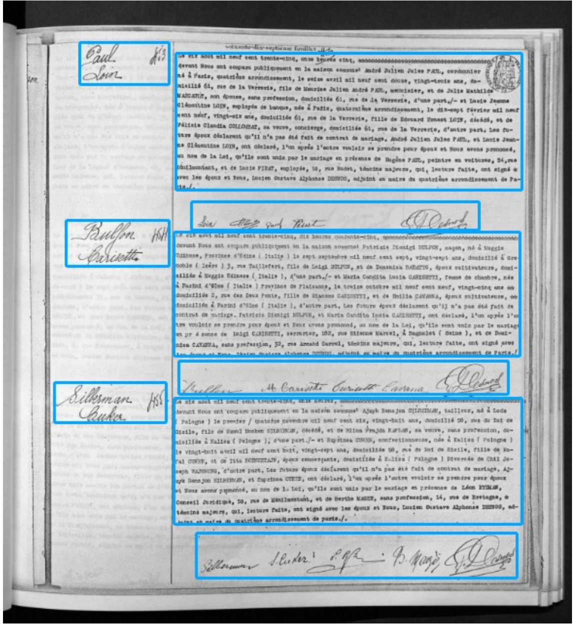
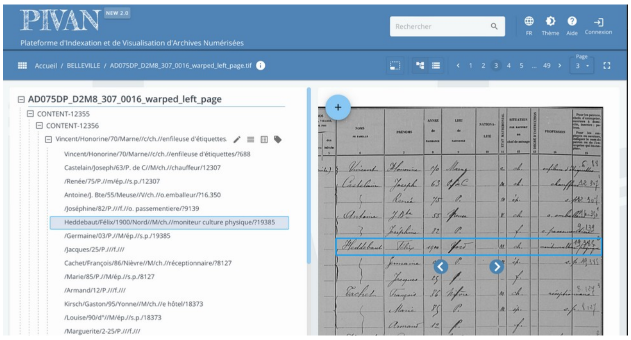


***Figure 2***. *An example of visualization of layout segmentation*

### Transcription

PIVAN allows side-by-side visualization and editing of a document image and its associated transcription. (See Figure 3) A user can start the transcription from scratch or edit the results of manual or automatic transcription. The transcription can be associated with a layout at the line or paragraph level. If segmentation is only available for paragraphs but not for lines, it is possible to enter the transcription of a paragraph in a single box and then the bounding box of each line will be automatically deduced by interpolation. Pivan includes a very useful feature: each layout element (page, paragraph, line) can be referenced by a specific URL. For example, in the case of a collaborative annotation campaign, this will allow listing in a shared spreadsheet a list of elements that need to be annotated or corrected, and each element will be directly accessible from the spreadsheet by clicking. This avoids wasting time going through different documents and paragraphs to find the item in question

.


PIVAN
NEW 2.0
Plateforme d'Indexation et de Visualisation d'Archives Numérisées
Rechercher
FR
Thème
Aide
Connexion
Page

Accueil / BELLEVILLE / AD075DP_D2M8_307_0016_warped_left_page.tif

⊟ AD075DP_D2M8_307_0016_warped_left_page
⊟ CONTENT-12355
⊟ CONTENT-12356
⊟ Vincent/Honorine/70/Marne//c/ch.//enfileuse d'étiquettes.
Vincent/Honorine/70/Marne//c/ch.//enfileuse d'étiquettes/?688
Castelain/Joseph/63/P. de C//M/ch.//chauffeur/12307
/Renée/75/P.//m/ép.//s.p./12307
Antoine/J. Bte/55/Meuse//V/ch.//o.emballeur/?16.350
/Joséphine/82/P.///f.//o. passementiere/?9139
Heddebaut/Félix/1900/Nord//M/ch.//moniteur culture physique/?19385
/Germaine/03/P.//M/ép.//s.p./19385
/Jacques/25/P.///f.///
Cachet/François/86/Nièvre//M/ch.//réceptionnaire/?8127
/Marie/85/P.//M/ép.//s.p./8127
/Armand/12/P.///f.///
Kirsch/Gaston/95/Yonne//M/ch.//e hôtel/18373
/Louise/90/d°//M/ép.//s.p./18373
/Marguerite/2-25/P.///f.///

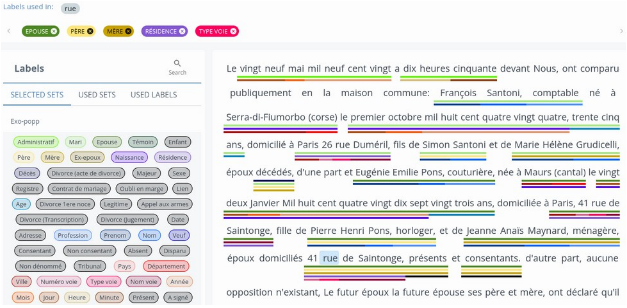
***Figure 3**. The interface for text annotation*

### *Named entity annotation*

PIVAN includes a tool to visualize and annotate named entities on transcribed documents. Unlike most other tools for this task, Pivan makes it possible to clearly visualize words with several named entity labels at the same time as shown in Figure 4.

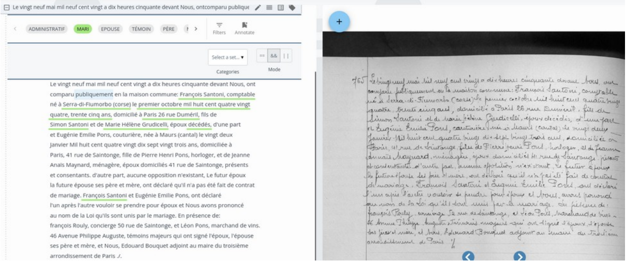
Pivan for named entities has two distinct interfaces. On the one hand, a visualization interface directly integrated into the rest of Pivan allows one to visualize the annotated named entities and to apply different filters to see only a part of them. On the other hand, an annotation interface is presented in the form of a dedicated page when the user chooses to annotate a given paragraph. This allows to display both the text to annotate and the annotation in progress more clearly.

It is possible to define multiple Named Entity labels set, either with the UI or via the API. For each named entity, it is possible to define a description, a color, and a URI. Named entity annotation can be very time-consuming. That is why Pivan for NER was made to be easy to use: with one action it is possible to annotate multiple words with the selected named entity labels. It is also possible to remove a given named entity label to several words at the same time. The annotation is made automatically at the token level. This means that the user does not need to select every character of a word to annotate it, but he simply has to double-click on a word or highlight with the mouse the words to annotate.



Le vingt neuf mai mil neuf cent vingt a dix heures cinquante devant Nous, ont comparu publiquement en la maison commune: François Santoni, comptable né à Serra-di-Fiumorbo (corse) le premier octobre mil huit cent quatre vingt quatre, trente cinq ans, domicilié à Paris 26 rue Duméril, fils de Simon Santoni et de Marie Hélène Grudicelli, époux décédés, d'une part et Eugénie Emilie Pons, couturière, née à Maurs (cantal) le vingt deux Janvier Mil huit cent quatre vingt dix sept vingt trois ans, domiciliée à Paris, 41 rue de Saintonge, fille de Pierre Henri Pons, horloger, et de Jeanne Anaïs Maynard, ménagère, époux domiciliés 41 rue de Saintonge, présents et consentants. d'autre part, aucune opposition n'existant, Le futur époux la future épouse ses père et mère, ont déclaré qu'il

***Figure 4**. The interface for named entity annotation*



Le vingt neuf mai mil neuf cent vingt a dix heures cinquante devant Nous, ontcomparu publique

Le vingt neuf mai mil neuf cent vingt a dix heures cinquante devant Nous, ont comparu publiquement en la maison commune: François Santoni, comptable né à Serra-di-Fiumorbo (corse) le premier octobre mil huit cent quatre vingt quatre, trente cinq ans, domicilié à Paris 26 rue Duméril, fils de Simon Santoni et de Marie Hélène Grudicelli, époux décédés, d'une part et Eugénie Emilie Pons, couturière, née à Maurs (cantal) le vingt deux Janvier Mil huit cent quatre vingt dix sept vingt trois ans, domiciliée à Paris, 41 rue de Saintonge, fille de Pierre Henri Pons, horloger, et de Jeanne Anaïs Maynard, ménagère, époux domiciliés 41 rue de Saintonge, présents et consentants. d'autre part, aucune opposition n'existant, Le futur époux la future épouse ses père et mère, ont déclaré qu'il n'a pas été fait de contrat de mariage. François Santoni et Eugénie Emilie Pons, ont déclaré l'un après l'autre vouloir se prendre pour époux et Nous avons prononcé au nom de la Loi qu'ils sont unis par le mariage. En présence de: françois Rouly, concierge 50 rue de Saintonge, et Léon Pons, marchand de vins. 46 Avenue Philippe Auguste, témoins majeurs qui ont signé l'époux, l'épouse ses père et mère, et Nous, Edouard Bouquet adjoint au maire du troisième arrondissement de Paris ./.

***Figure 5**. The interface for named entity visualization.*

## Technology

It is possible to deploy the user's own Pivan instance online or offline to host the user's project images and annotations. The encapsulation of Pivan in a Docker image makes it easy to deploy, even for people with no knowledge of networking or web development. Because of its web-based nature, Pivan is perfectly suited for collaborative annotation. It is free of operating system requirements for its users but must be set up by someone with some Web deployment skills.

Pivan is based on modern web technologies with the JavaScript framework React on the frontend and the Java framework Spring Boot on the backend. The two are connected by a RESTful API based on Swagger. The technologies behind Pivan allow this platform to offer a smooth experience, even for small machines or connections. In addition, the platform has been designed to be responsive for different screen sizes, including tablet screens. More details regarding the architecture can be found in Figure 6.

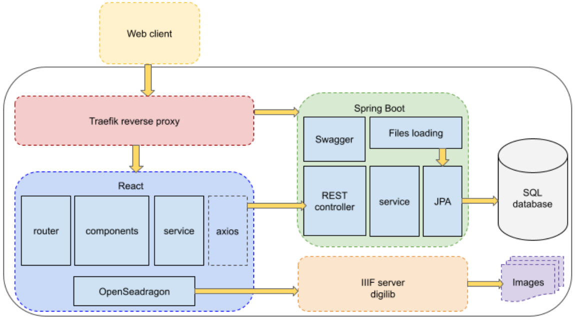


***Figure 6**. Diagram of the architecture of Pivan*

### API

Pivan has a comprehensive API with endpoints for every operation that can be performed via the user interface such as managing users and their access rights, annotating text and named entities, and export collections. For example, it is possible to annotate named entities automatically via a script using regular expressions which would perform the annotation via the API.

### Import and export of collections

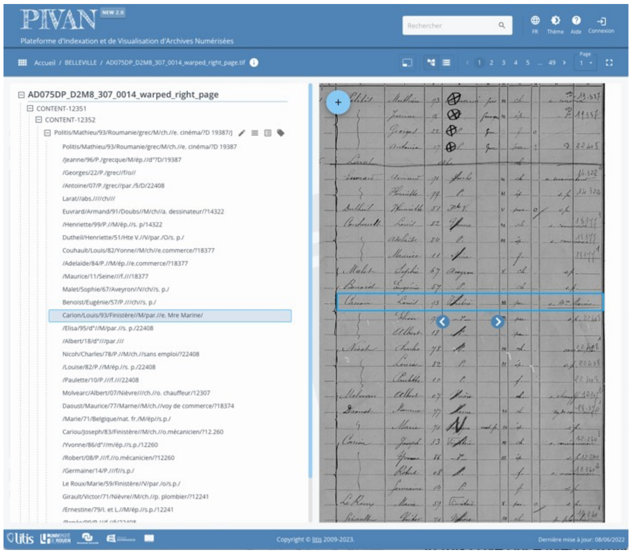
Pivan allows importing a collection in METS/ALTO format. The collection may or may not contain transcription annotations and named entities annotations. However, the files must contain paragraph segmentation since bounding box editing is not currently available.

The API allows to export an entire collection in METS/ALTO format, or to export a specific element of a collection (annotation of a page, line, or word) which will be received as a JSON payload that can then be exploited directly by a script.

## Use-cases

### Early 20th-century Paris census

During the POPP project [2], we built an HTR system to extract population information from handwritten censuses (Figure 7).


PIVAN NEW 2.0
Plateforme d'Indexation et de Visualisation d'Archives Numérisées
Rechercher
FR
Thème
Aide
Connexion
Accueil / BELLEVILLE / AD075DP_D2M8_307_0014_warped_right_page.tif
1 2 3 4 5 ... 49
Page
1
Copyright © litis 2009-2023
Dernière mise à jour: 08/06/2022

AD075DP_D2M8_307_0014_warped_right_page
CONTENT-12351
CONTENT-12352
Politis/Mathieu/93/Roumanie/grec/M/ch.//e. cinéma/?D 19387/J
Politis/Mathieu/93/Roumanie/grec/M/ch.//e. cinéma/?D 19387
/Jeanne/96/P./grecque/M/ép.//d°?D/19387
/Georges/22/P./grec//f/o//
/Antoine/07/P./grec//par./f/D/22408
Larat//abs.////ch///
Euvrard/Armand/91/Doubs//M/ch//a. dessinateur/?14322
/Henriette/99/P.//M/ép.//s. p/14322
Dutheil/Henriette/51/Hte V.//V/par./O/s. p./
Couhault/Louis/82/Yonne//M/ch//e.commerce/?18377
/Adelaide/84/P.//M/ép.//e.commerce/?18377
/Maurice/11/Seine///f.///18377
Malet/Sophie/67/Aveyron//V/ch//s. p./
Benoist/Eugénie/57/P.///ch//s. p./
Carion/Louis/93/Finistère//M/par.//e. Mre Marine/
/Elisa/95/d°//M/par.//s. p./22408
/Albert/18/d°///par.///
Nicoh/Charles/78/P.//M/ch.//sans emploi/?22408
/Louise/82/P.//M/ép.//s. p./22408
/Paulette/10/P.///f.///22408
Molvearc/Albert/07/Nièvre///ch.//o. chauffeur/12307
Daoust/Maurice/77/Marne//M/ch.//voy de commerce/?18374
/Marie/71/Belgique/nat. fr./M/ép//s.p./
Cariou/Joseph/83/Finistère//M/ch.//o.mécanicien/?12.260
/Yvonne/86/d°//m/ép.//s.p./12260
/Robert/08/P.///f.//o.mécanicien/?12260
/Germaine/14/P.///f//s.p./
Le Roux/Marie/59/Finistère//V/par./o/s.p./
Girault/Victor/71/Nièvre//M/ch.//p. plombier/?12241
/Ernestine/79/I. et L.//M/ép.//s.p./12241

*Figure 7. A line from the census and the corresponding text in PIVAN*

After an automatic page and line segmentation, we used PIVAN to produce the transcription ground truth upon which our system could be trained and tested later on.

### 1880 - 1930 Marriage records

In the EXOPOPP project [4] we aim at extracting all the relevant population information from marriage records, from the birthplace of the groom to the job title of the bride's mother or the home address of a witness. We used PIVAN to produce the ground truth for our layout and segmentation-free HTR system (DAN [5]), and then visualize and correct the results. We treated each bit of information as a named entity tag, as illustrated in Figure 4, and used PIVAN to build the ground truth for our system to be trained upon. Digital document specialists Numen [6] used PIVAN to create and correct another corpus of ground truth.

### Demonstrator

A working demonstrator of PIVAN can be accessed at litis-pivan.univ-rouen.fr. It can be used to peruse different collections, but not modify transcriptions or named-entities: these functionalities are restricted for security reasons on the open demonstrator.

The open access to the software should be ready before the end of 2023. A link to it will then be added in the demonstrator website.

## Conclusion

To conclude, Pivan is an open-source web platform that simplifies and accelerates the process of document annotation for different stages of automatic document processing. It provides efficient tools for layout analysis, transcription, and named entity recognition, making it a versatile and user-friendly platform. Pivan's web-based nature and RESTful API make it an ideal tool for collaborative annotation and customization. With its support for METS/ALTO format and export capabilities, Pivan can easily be integrated into existing document processing workflows. At LITIS, we used Pivan for several projects such as information extraction of census tables of Paris [2], OCR applied on stock market prices [3] and handwriting recognition of Norman law manuscripts.

## References

[1] P Tranouez, S Nicolas, J Lerouge, T Paquet. PIVAJ: an article-centered platform for digitized newspapers, Archiving Conference 2015, 40-43, Los Angeles, 05/2015

[2] T. Constum, N. Kempf, T. Paquet, P. Tranouez, C. Chatelain, et al , Recognition and Information Extraction in Historical Handwritten Tables : Toward Understanding Early 20th Century Paris Census. Document Analysis Systems 2022 proceedings, 13237, Springer International Publishing, pp.143-157, 2022, Lecture Notes in Computer Science

[3] Sébastien Adam, Simon Bouvier, Bertrand B. Coüasnon, Nathalie Girard, Camille Guerry, et al. EURHISFIRM - M7.2: Final version of the data extraction system. [Research Report] European Union's Horizon 2020 research and innovation programme. 2021.

[4] https://exopopp.hypotheses.org/

[5] D. Coquenet, C. Chatelain and T. Paquet, "DAN: a Segmentation-free Document Attention Network for Handwritten Document Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, doi: 10.1109/TPAMI.2023.3235826.

[6] https://www.numen.fr/qui-sommes-nous

## Authors Biography

***Thomas Constum*** *received the Engineering Degree in Information Systems Architecture at the National Institute of Applied Sciences in Rouen, France. He is a second-year Ph.D. student at the University of Rouen Normandy. His research interests include deep* *learning approaches, handwriting text recognition and named entity recognition.*

***Florian Bebin*** *was a research engineer at LITIS in charge of the development of Pivan. He is now a software development engineer at the company SII.*

***Pierrick Tranouez*** *is Research Engineer at the University of Rouen Normandy. His research interests are agent-based modeling and simulation of complex systems, and machine learning applied to document image analysis and historical documents.*

***Thierry Paquet*** *is Professor at the University of Rouen Normandy. His research interests are machine learning, statistical pattern recognition, deep learning, sequence modeling, with applications to document image analysis and handwriting recognition.*