

The FINLAM project : *Outlining the State of the Art in Newspapers Segmentation*

ONB Labs Symposium 2024
Newspapers as Datasets

<https://finlam.projets.litislabs.fr/>

Overview

1. What is the FINLAM project, who's involved ?
2. Challenges of digitizing newspapers at the BnF
3. Technical options being considered (for now)

1. What's the FINLAM project, who's involved ?

Foundation **I**Ntegrated models for **L**ibraries, **A**rchives and **M**useums

Funded by the **N**ational **A**gency of **R**esearch - 4 years - started in mid-2023.

- Aims to exploit **LLMs** to build new **multimodal** models combining vision and language to process heritage documents in **heritage institutions**.
- Aims to support new modes of **interaction** with documents, enabling question answering about document structure, named entities or metadata extraction.
- Aims to **reduce the distance** between end-users (curators) and their documents by allowing on-the-fly multimodal data extraction.



1. What's the FINLAM project, who's involved ?

Partners :

The LITIS lab (university of Rouen)

- Specialised in Statistical Pattern Recognition and Deep Neural Networks
- with application to document images processing and text analysis and recognition.

TEKLIA

- French software publisher.
- Develops innovative solutions for semantic processing of digitized documents (machine learning, deep learning, and NLP).
- Has established itself as a frontrunner in the automatic processing of historical and heritage documents.

1. What's the FINLAM project, who's involved ?

Implementation of the R&D approach as part of the schedule and budget

	Year 1				Year 2				Year 3				Year 4			Human Resources
	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15	
WP0 Management								M0.1							M0.2	8 MM
WP1 Data				M1.1			M1.2									18 MM
WP2 Models					M2.1				M2.2			M2.3				80 MM
WP3 Applications									M3.1				M3.2			55 MM
WP4 Industrialization												M4.1			M4.2	38 MM

First dataset available on HF (150 issues, ~700 pages) :

<https://huggingface.co/datasets/Teklia/Newspapers-finlam/viewer>

2. Challenges of digitizing newspapers at the BnF

2029 : opening of the [National Press Conservatory](#), in Amiens.

Amiens



- Will bring together BnF's press massive collections - more than [270,000 titles](#).
- 2,000 m² for restoration and digitization workshops.
- 6,000 m² storage area with a capacity of 280 linear kilometers in an oxygen-depleted atmosphere.
- [National plan for Press digitization](#) (15 million pages, 18 million €).



2. Challenges of digitizing newspapers at the BnF

2016-2019 : first public contract for a press enrichment service

- 233,241 issues (1,223,000 pages) of national and international, large-format, royalty-free press (late 19th > mid-20th century).
- National and European funds (~1.25 € / page for digitization, OCR and OLR).
- Production by human operators.

- Technical specifications :
 - article segmentation ;
 - categorization (headings, writers names, advertising, ads, novels and legal / financial chronicles).



2. Challenges of digitizing newspapers at the BnF

Expected accuracy	
Accuracy rate for text transcription	No minimum rate for articles content, but for headings and writers names
Accuracy rate for articles segmentation, and categorization	98,5 %

Expected accuracy	
Identification of article headings (title, subtitle), inner headings and writers names	95 %
Transcription	99,5 %

3. Technical options being considered

An first set of newspapers has been uploaded in ArkIndex (<https://doc.arkindex.org/>) :

- 1,500 issues from the above-mentioned set.
- XML METS / ALTO files converted in JSON.
- Usage of the JSON :
 - end-to-end model input/output ;
 - evaluation library input.



3. Technical options being considered

Available annotations in ArkIndex :

- Text content (varying quality).
- Zones coordinates.
- Zones labeling :
 - textblock, illustration...
 - title, subtitle, advertisement, freead...
- Complete structural and logical maps for each issue.

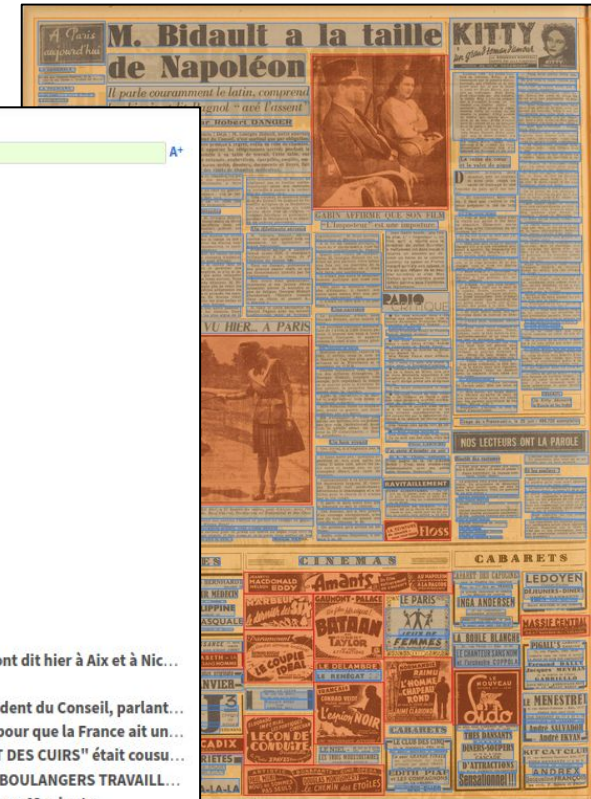
« Chaque fois que le Gouvernement fera fusiller l'un des nôtres, déclare le général en chef des troupes rebelles, nous fusillerons deux gouvernementaux »

« Chaque fois que le Gouvernement fera fusiller l'un des nôtres, déclare le général en chef des troupes rebelles, nous fusillerons deux gouvernementaux »

Filter by type...

NEWSPAPER DIV.8 A*

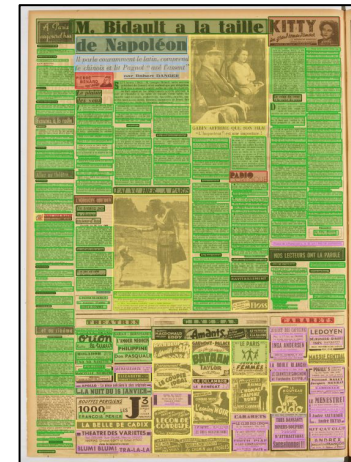
- ▼ ISSUE DIV.9
 - ▼ TITLESECTION DIV.10
 - HEADLINE DIV.11
 - TEXTBLOCK DIV.12
 - TEXTBLOCK DIV.13
 - TEXTBLOCK DIV.14
 - TEXTBLOCK DIV.15
 - TEXTBLOCK DIV.16
 - TEXTBLOCK DIV.17
 - TEXTBLOCK DIV.18
 - TEXTBLOCK DIV.19
 - TEXTBLOCK DIV.20
 - TEXTBLOCK DIV.21
 - TEXTBLOCK DIV.22
 - TEXTBLOCK DIV.23
 - ▼ CONTENT DIV.24
 - ▼ ARTICLE Annonces
 - BODY DIV.26
 - ▼ ARTICLE Annonces
 - BODY DIV.30
 - ▼ SECTION "VOTEZ OUI" ont dit hier à Aix et à Nic...
 - HEADING DIV.37
 - ARTICLE Le vice-président du Conseil, parlant...
 - ARTICLE Votez "oui" pour que la France ait un...
 - ARTICLE LE "PRESIDENT DES CUIRS" était cousu...
 - ARTICLE LES OUVRIERS BOULANGERS TRAVAIL...
 - ARTICLE Paris-Bruxelles en 43 minutes
 - ARTICLE Une caissette cachée en pays de Bade p...
 - ARTICLE A Longchamp : M. Byrnes A JOUÉ ET GA...
 - ARTICLE LES " 4 " SE RÉUNISSENT A 16 H. : Le ré...
 - ARTICLE LE VIEUX VITRÉ PARTIELLEMENT DÉTR...
 - ARTICLE ÉCOLES ET LYCÉES VAQUERONT LE 1er ...
 - ARTICLE « ANDERS ARME LES BANDES FASCISTE...



3. Technical options being considered

FINLAM's tasks on newspapers

	Sub-tasks	Difficulties
Page layout	<ul style="list-style-type: none"> - Detection of text/illustration blocks - Zone classification - Headings tree 	<ul style="list-style-type: none"> - Complex layout - Confusion between section or article / title or subtitle
Text recognition	<ul style="list-style-type: none"> - Recognition - Post-OCR processings 	<ul style="list-style-type: none"> - Physical state of the document - Text density
Articles segmentation	<ul style="list-style-type: none"> - Locate separators - Gather right paragraphs into articles 	<ul style="list-style-type: none"> - Multi-page articles
Reading order		<ul style="list-style-type: none"> - Complex layout



3. Technical options being considered





Available datasets

- OCR : <https://zenodo.org/records/4293602> / <https://zenodo.org/records/4599624> / <https://zenodo.org/records/4599472>
- Post OCR correction: <https://zenodo.org/records/3515403>
- Article separation : <https://zenodo.org/records/5654841> / <https://zenodo.org/records/5654907> / <https://zenodo.org/records/5654858>
- Segmentation : <https://zenodo.org/records/4943582>

- DIY History : <https://diyhistory.lib.uiowa.edu/>
- Europeana Transcribathon : <https://www.europeana.eu/en>
- HTR-United : <https://htr-unity.github.io/>
- IMPACT : <https://lab.kb.nl/dataset/ground-truth-impact-project>

3. Technical options being considered

Available models

- Arcanum
 - Layout 
 - Text 
 - Reading order 
 - Articles 
- Efficient but chargeable tool



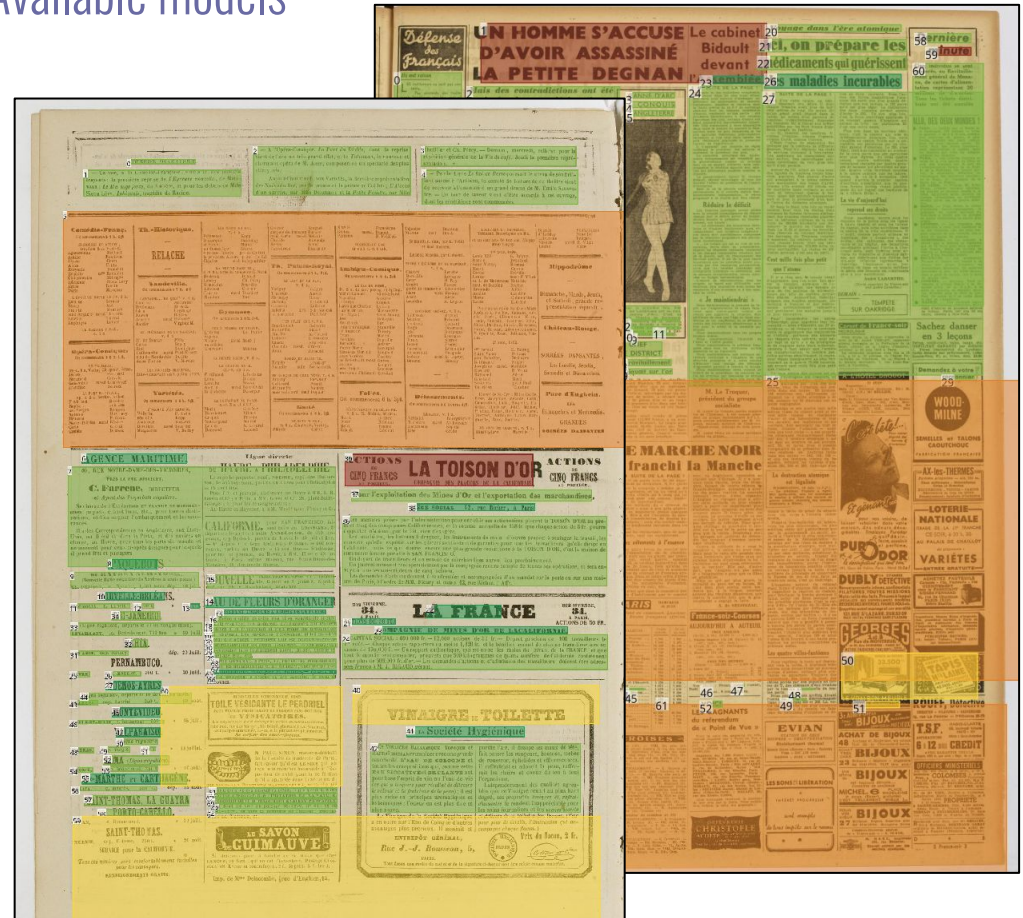
3. Technical options being considered

Available models

- Surya

- Layout ✓
- Text ✓
- Reading order ✓
- Articles ✗

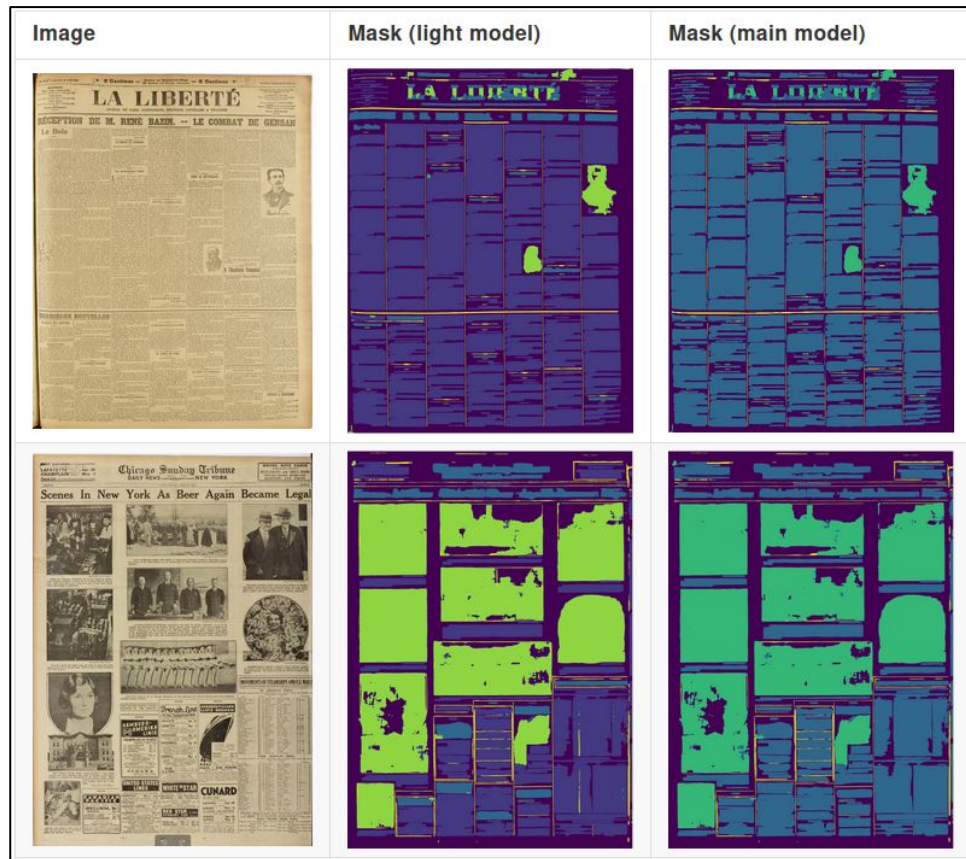
- Open-source
- Need to be fine-tuned



3. Technical options being considered





Available models

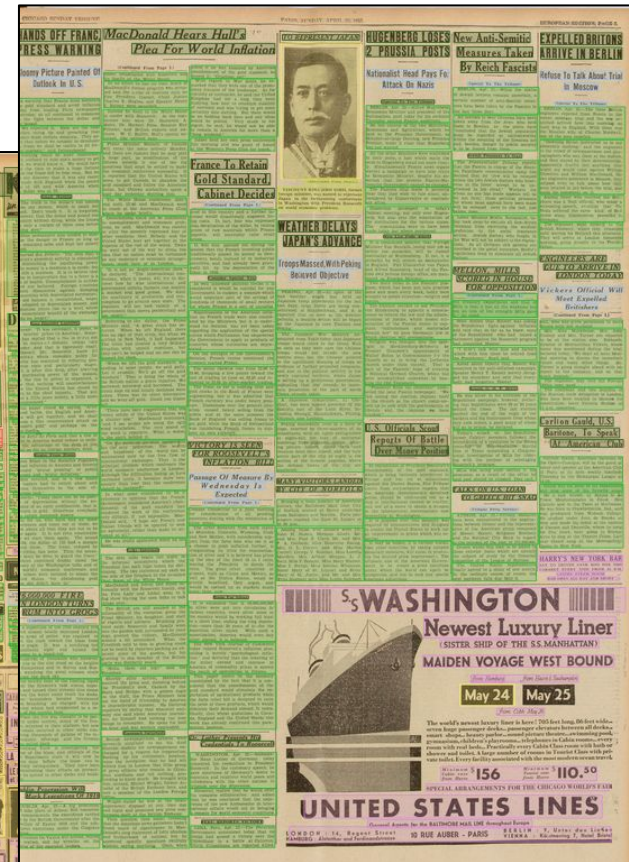
- Eynollah
 - Layout ✓
 - Text ✗
 - Reading order ✗
 - Articles ✗
- Open-source
- Need to be fine-tuned



3. Technical options being considered


Available models

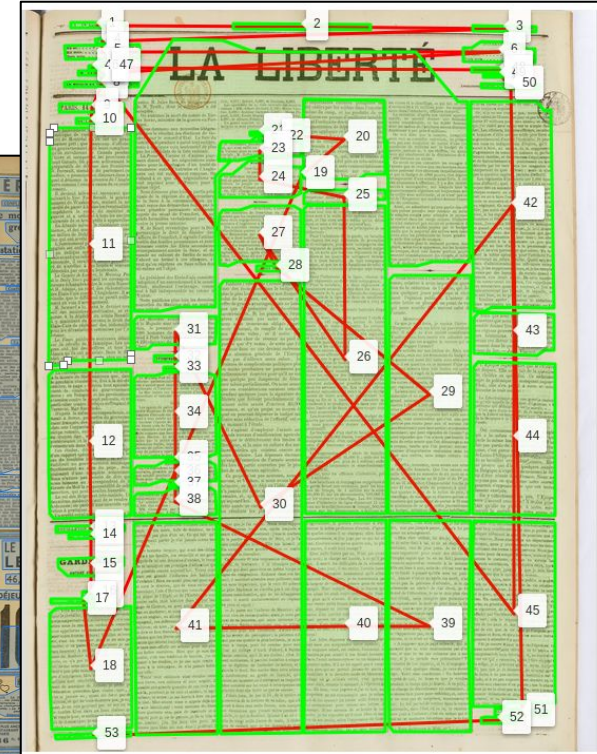
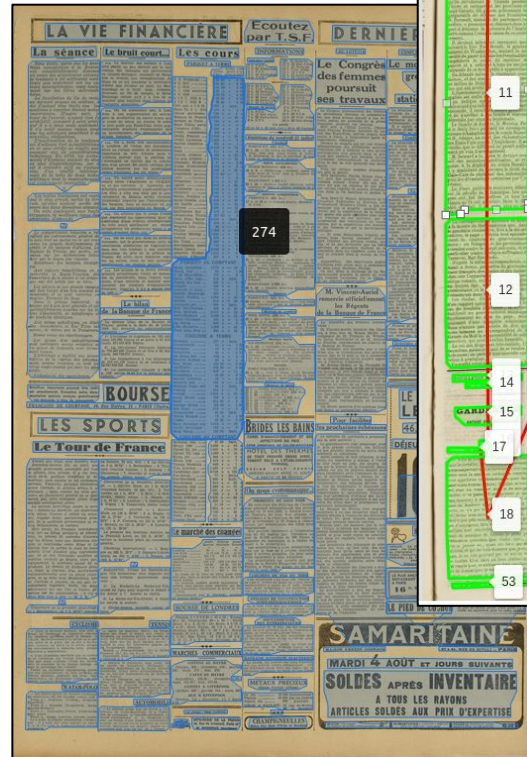
- YOLO
 - Layout 
 - Text 
 - Reading order 
 - Articles 
- Open-source
- Need to be fine-tuned



3. Technical options being considered

Available models

- PERO-OCR
 - Layout 
 - Text 
 - Reading order 
 - Articles 
- Open-source





3. Technical options being considered

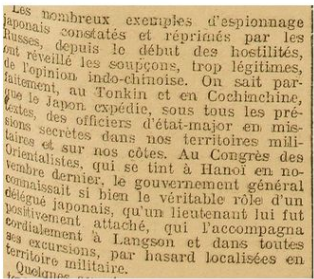
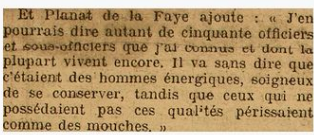
Available models

*“You're an OCR system. Transcribe this entire **paragraph** from a French newspaper, keep the reading order. Format the output in Markdown, keep the headings structure.”*

- MLLMs ([QWEN2](#))

- Mise en page 
- Texte  (paragraphe )
- Ordre de lecture 
- Articles 

- Open-source

Image	Prediction
	Les nombreux exemples d'espionnage japonais constatés et réprimés par les Russes, depuis le début des hostilités, ont réveillé les soupçons, trop légitimes, de l'opinion indo-chinoise. On sait par exemple, au Tonkin et en Cochinchine, que le Japon expédie, sous tous les prétextes, des officiers d'état-major en missions secrètes dans nos territoires et sur nos côtes. Au Congrès des vétérans, qui se tint à Hanoï en novembre dernier, le gouvernement général délégué le véritable rôle d'un positivement attaché, qui l'accompagna dans ses excursions, par hasard localisées en quelques cas, dans le Tonkin et en Cochinchine.
	Et Planat de la Faye ajoute : « J'en pourrais dire autant de cinquante officiers et sous-officiers que j'ai connus et dont la plupart vivent encore. Il va sans dire que c'étaient des hommes énergiques, soigneux de se conserver, tandis que ceux qui ne possédaient pas ces qualités périssaient comme des mouches. »

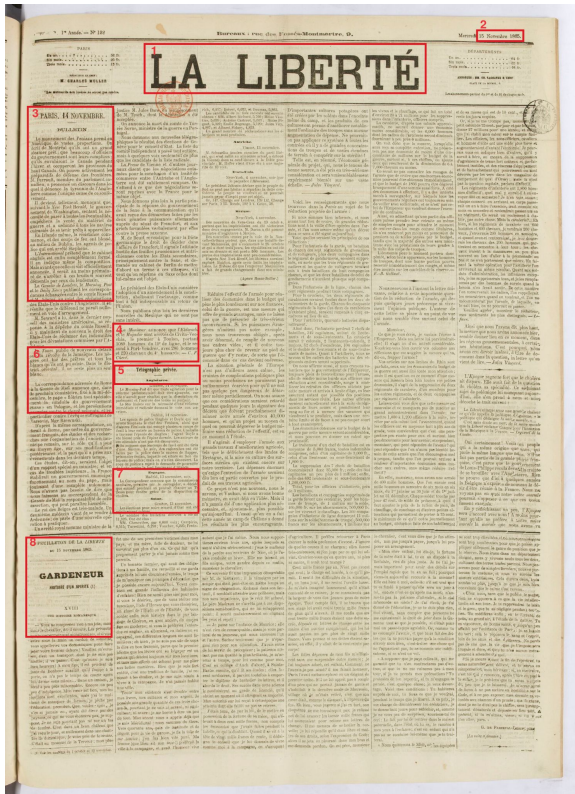
*“You’re an OCR system. Transcribe this entire **page** from a French newspaper, keep the reading order. Format the output in Markdown, keep the headings structure.”*

[illegible]

3. Technical options being considered

Available models

MLLMs (QWEN2)



*“Extract only the headings on this page from a French newspaper.
Keep the reading order of the page.”*

1. ****Bureaux : rue des Fossés-Montmartre, 9.****
2. ****Mercredi 15 Novembre 1865.****
3. ****PARIS, 14 NOVEMBRE.****
4. ****BULLETIN****
5. ****PARIS, 14 NOVEMBRE.****
6. ****GARDENIER****
7. ****HISTOIRE D'UN SPIRITÉ (1)****
8. ****XVIII****
9. ****UNE HISTOIRE ROMANESQUE****
10. ****FEUILLETON DE LA LIBERTÉ****
11. ****DU 15 NOVEMBRE 1865.****
12. ****GARDENIER****
13. ****HISTOIRE D'UN SPIRITÉ (1)****
14. ****XVIII****
15. ****UNE HISTOIRE ROMANESQUE****
16. ****FEUILLETON DE LA LIBERTÉ****
17. ****DU 15 NOVEMBRE 1865.****
18. ****GARDENIER****

3. Technical options being considered

Available models - not tested yet

- [LIAS](#) : Layout Information-Based Article Separation in Historical Newspapers.
- [STRAS](#) : A Semantic Textual-Cues Leveraged Rule-Based Approach for Article Separation in Historical Newspapers.
- [LayoutReader](#)

- # Saint-Christophe-du-Foc
- ## Saint-Christophe du Foc est une commune française.

Géographie

Climat

Le climat qui caractérise la commune est qualifié, en 2010, de « climat océanique franc », selon la typologie des climats de la France qui compte alors huit grands types de climats en métropole. En 2020, la commune ressort du type « climat océanique » dans la classification établie par Météo-France, qui ne compte désormais, en première approche, que cinq grands types de climats en métropole. Ce type de climat se traduit par des températures douces et une pluviométrie relativement abondante (en liaison avec les perturbations venant de l'Atlantique), répartie tout au long de l'année avec un léger maximum d'octobre à février.

Urbanisme

Typologie

Saint-Christophe-du-Foc est une commune rurale, car elle fait partie des communes peu ou très peu denses, au sens de la grille communale de densité de l'Insee.

Par ailleurs la commune fait partie de l'aire d'attraction de Carentan, dont elle est une commune de la couronne. Cette aire, qui regroupe, est caractérisée dans les années 2010 par la faible densité.

L'occupation des sols

L'occupation des sols de la commune, telle qu'elle ressort de la base de données européenne d'occupation géophysique des sols Corine Land Cover (CLC), est marquée par l'importance des territoires agricoles (100 % en 2018), une proportion identique à celle de 1990 (100 %). La répartition détaillée en 2018 est la suivante :

 - terres arables (81 %)
 - prairies non irriguées (19 %)
 - terres non agricoles (6,6 %)

Toponymie

Le nom du lieu est attesté sous la forme *Sanctus Christophorus* vers 1000, *Fagum* 1066 - 1083, *Fagum* 1080 - 1082, *Sancti Christophori* vers 1100 et 1281 ainsi que sous la forme *Sancti Christophori* d' *Fagum* en 1468. Gilles de Gouberville le désigne dans les années 1580 comme *Sancti Christophori de Fagum* et *Christophori de la Chapelle*.

Le nom du lieu

Le mot *Sanctus* est le latin pour *Saint*. *Christophorus* est le latin pour *Christ*. *Fagum* est le latin pour *forêt*. Le nom du lieu est donc *Sanctus Christophorus de Fagum*.

Le nom du lieu

Le mot *Sanctus* est le latin pour *Saint*. *Christophorus* est le latin pour *Christ*. *Fagum* est le latin pour *forêt*. Le nom du lieu est donc *Sanctus Christophorus de Fagum*.

Le nom du lieu

Le mot *Sanctus* est le latin pour *Saint*. *Christophorus* est le latin pour *Christ*. *Fagum* est le latin pour *forêt*. Le nom du lieu est donc *Sanctus Christophorus de Fagum*.

Le nom du lieu

Le mot *Sanctus* est le latin pour *Saint*. *Christophorus* est le latin pour *Christ*. *Fagum* est le latin pour *forêt*. Le nom du lieu est donc *Sanctus Christophorus de Fagum*.

Thanks for your attention !

Contact : sebastien.cretin@bnf.fr