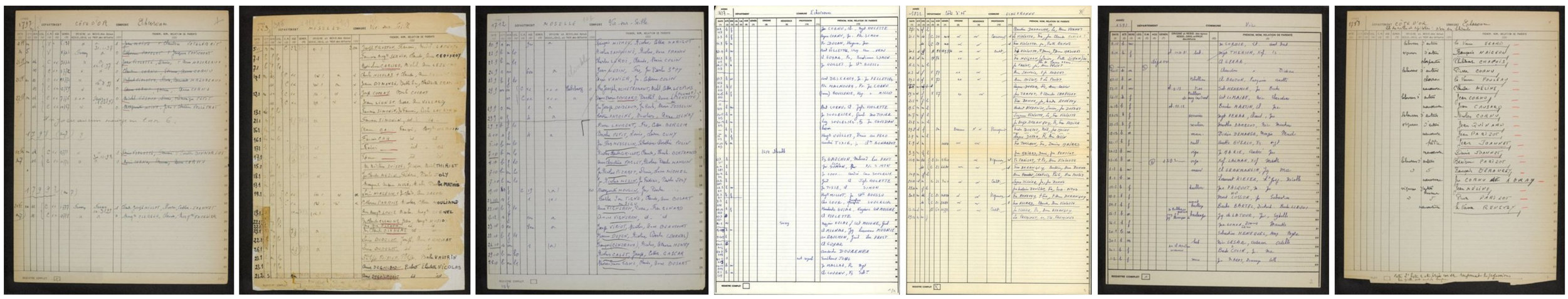
Un corpus de tables de recensement historiques : publication, analyse et extraction de lignes

Guillaume Bernard, Casey Wall, Mickaël Coustaty, Antoine Doucet

Un corpus de tables de recensement historiques
Publication, analyse et extraction de lignes

Guillaume Bernard, Casey Wall, Mickaël Coustaty, Antoine Doucet
Laboratoire L3i, Université de La Rochelle



Contexte

- La **numérisation des contenus** des **archives** (bibliothèques, archives départementales, nationales) est un enjeu contemporain.
- Les **tables de recensement** peuvent être utilisés pour les **études démographiques** : diffusion de noms de famille en relation avec certains événements connus du passé. Par exemple, la Première Guerre Mondiale a mené à l'exode vers l'ouest de la France de population de Somme et de la Marne, entre autres.
- Des milliers de tables de recensement numérisées : ce sont des **images de documents manuscrits**. Pour 537 documents, il existe une **transcription complète**.

Enjeux scientifiques

- **Reconnaissance de mise en page** : détection et extraction des constituants de la page : lignes de texte, en-têtes.
- **Extraction du texte manuscrit** : reconnaissance du texte de chaque ligne, avec des niveaux de dégradation variés.
- Utilisable pour **traiter des données en volume** : vitesse et coût énergétique.

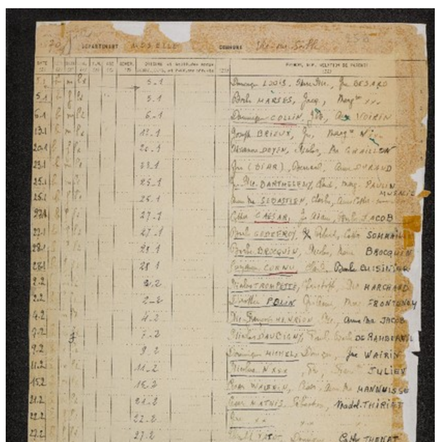
Jeu de données

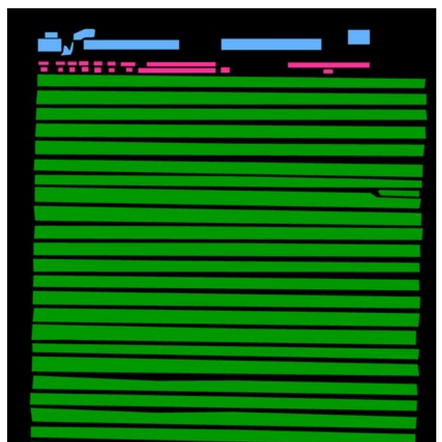
doi 10.5281/zenodo.7840570

- Nous **diffusions** un échantillon de 250 documents, annotés pour la détection de mise en page.
- Répartis en 7 catégories (*cf* l'image en haut du document).

Catégorie	Description	Ratio
1	Aucun dégât visible	59,6 %
2	Fortes dégradations	0.8 %
3	Variantes de C1 avec différents fonds.	29,6 %
4	Mise en page différente, fond clair.	3,6 %
5	Variante jaunie de C4.	3,6 %
6	En tête différents de C1	0,4 %
7	Ne suite pas la mise en page du document.	2,4 %

Annotations



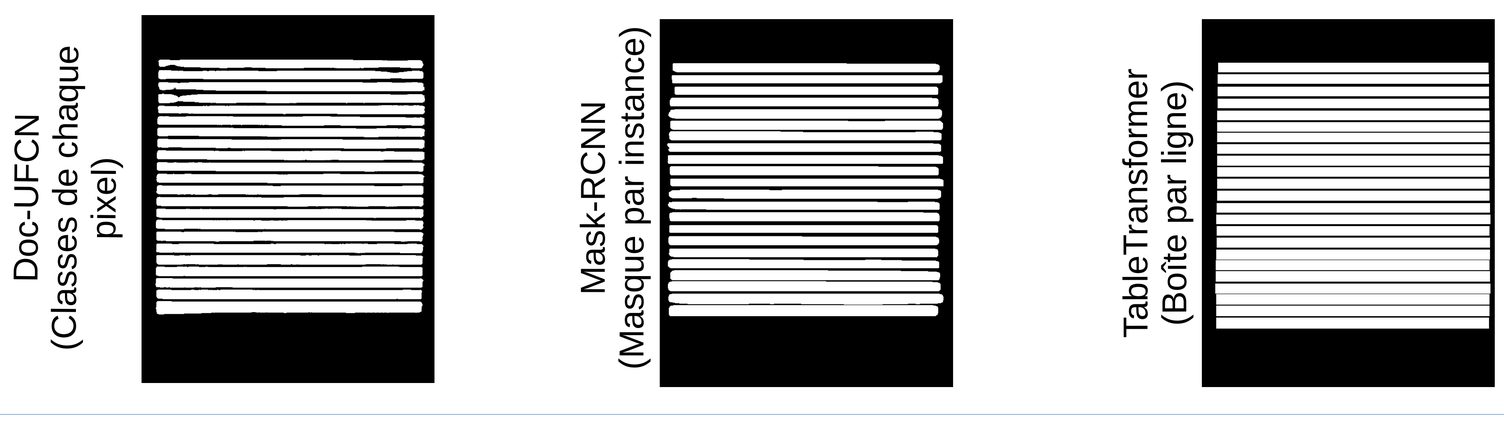


- En-tête de page
- Bas de page
- En-tête de tableau
- Ligne de texte

Image originale et annotation de ses différents composants

Expériences : détection de lignes de texte

1. Comparaison de 3 méthodes: **U-Nets**, **Mask-RCNN** et **Transformers**.
2. Entraînement sur un jeu de données stratifié :
   - `train` sur les catégories C1 et C2 pour évaluer la généralisation ;
   - `val` sur un mélange de C1 à C7 ;
   - `test` sur les documents les plus dégradés, un mélange de C1 à C7. Nous évaluons la capacité du modèle à détecter des lignes dans des documents fortement abimés ou inconnus.
3. Métriques d'*IR* de pixels (F1, Précision, Rappel) et d'objets (AP, AP[.5], AP[.75]).
4. Résultats comparables selon les modèles. `TableTransformer` a un meilleur FPS.



Travaux et défis futurs

- Concevoir ou réutiliser une architecture de ***transformer*** pour réaliser en une seule passe la **reconnaissance de la mise en page** ET l'**extraction du texte manuscrit**.
- Fonctionner avec **un nombre réduit d'annotations** pour l'entraînement.

• **Jeu de données sur Zenodo : https://doi.org/10.5281/zenodo.7840570**
• Boillet, Mélodie, Christopher Kermorvant, et Thierry Paquet. 2022. « Robust Text Line Detection in Historical Documents: Learning and Evaluation Methods ». International Journal on Document Analysis and Recognition (IJDAR) 25 (2): 95-114. https://doi.org/10.1007/s10032-022-00395-7.
• Droby, Ahmad, Berat Kurar Barakat, Reem Alaasam, Boraq Madi, Irina Rabaev, et Jihad El-Sana. 2022. « Text Line Extraction in Historical Documents Using Mask R-CNN ». Signals 3 (3): 535-49. https://doi.org/10.3390/signals3030032.
• Smock, Brandon, Rohith Pesala, et Robin Abraham. 2022. « PubTables-1M: Towards Comprehensive Table Extraction From Unstructured Documents ». In , 4634-42.

Réalisation pour le Symposium International Francophone sur l'Écrit et le Document (SIFED 2023) – 9 juin 2023 – Paris, France