HTR-United

Un catalogue pour les données d'entrainement à la reconnaissance de texte

https://htr-united.github.io

Thibault Clérice (ALMAnaCH)

Alix Chagué (ALMAnaCH, UdeM, ÉPHE)

28 Mars 2024

Webinaire pour le GTSO Données du consortium Couperin

Table of Contents

1 HTR : définition

- HTR : définition
- Situation
- Qu'est-ce que HTR-United?
- Enjeux et résultats
- Conclusion

Qu'est-que l'HTR ?

1 HTR : définition

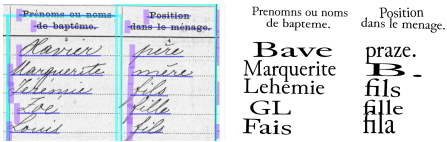


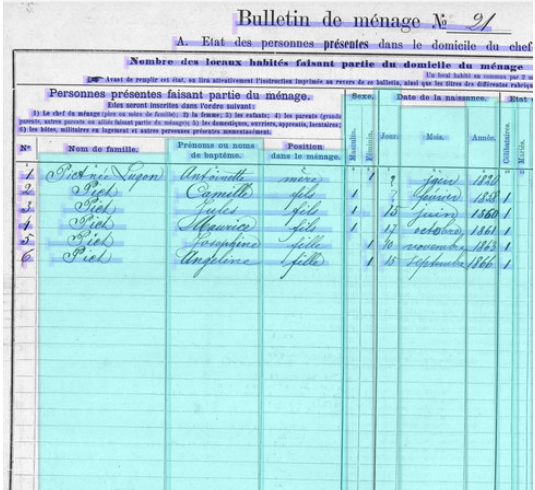
Figure: Prédiction HTR

- Prédiction d'un contenu texte
- à partir d'une image de la source par une
- intelligence artificielle entraînée par un humain
- dans un processus alternant
  - phases d'interventions humaines
  - et phases de calcul

Les étapes: Obtenir les images et interpréter la mise en page

1 HTR : définition

- Chargement des images
- Traitement des images (facultatif)
- Segmentation des zones de l'image



Bulletin de ménage N° 21

A. Etat des personnes présentes dans le domicile du chef

Nombre des locaux habités faisant partie du domicile du ménage.

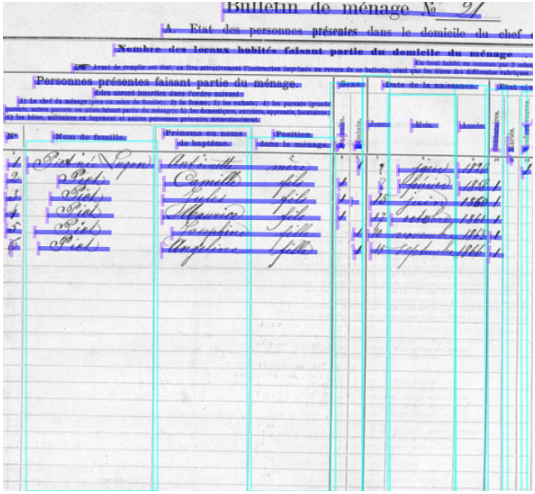
Personnes présentes faisant partie du ménage.

Sexe. Date de la naissance. Etat

N°	Nom de famille.	Prénoms ou noms de baptême.	Position dans le ménage.	Masculin.	Féminin.	Jour.	Mois.	Année.	Célibataires.	Mariés.
1	Pichonnier Lupon	Antoinette	mère		1	?	?	1818	"	"
2	Piel	Camille	fils	1		7	Janvier	1838	1	
3	Piel	Jules	fils	1		15	Janvier	1850	1	
4	Piel	Maurice	fils	1		17	octobre	1861	1	
5	Piel	Joséphine	fille		1	10	novembre	1843	1	
6	Piel	Angeline	fille		1	15	Septembre	1866	1	

Les étapes: trouver les lignes

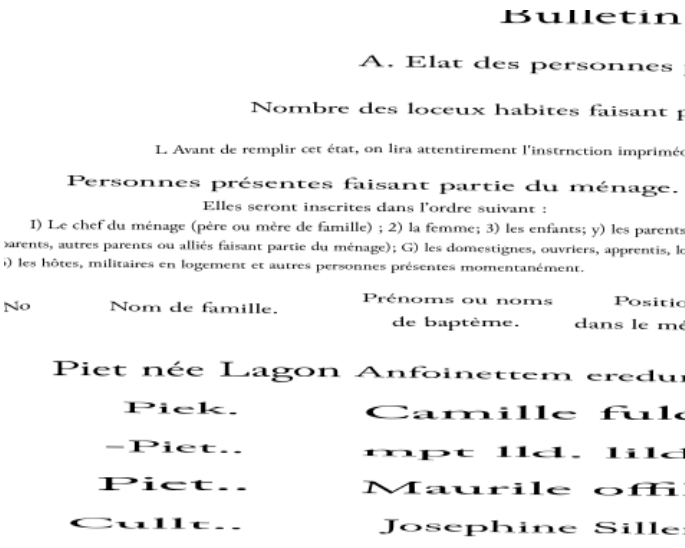
- Segmentation des lignes contenant du texte



Reconnaître le texte

1 HTR : définition

- Prédiction du texte qui se trouve sur les lignes



Rendre exploitable

1 HTR : définition

- Export des données (txt, alto, page, json)

```
<Layout>
        <Page WIDTH="4648" HEIGHT="3407" PHYSICAL_IMG_NR="8" ID="eSc_dummypage_">
                <PrintSpace HPOS="0" VPOS="0" WIDTH="4648" HEIGHT="3407">

                        <TextBlock HPOS="693" VPOS="321" WIDTH="1701" HEIGHT="2451
                                ID="eSc_textblock_08b9f915" TAGREFS="BT3852">
                                <Shape>
                                        <Polygon
                                          POINTS="693 413 693 2772 2394 2772 2254
                                </Shape>


                                <TextLine ID="eSc_line_d939596f" TAGREFS="LT1299"
                                        BASELINE="746 476 2143 428" HPOS="743" VPO
                                        WIDTH="1400" HEIGHT="156">
                                        <Shape>
                                          <Polygon
                                          POINTS="2078 388 2050 388 2021 386 1993
                                          />
                                        </Shape>
                                        <String
                                          CONTENT="fors de la uille. Tant fut lass
                                          HPOS="743" VPOS="352" WIDTH="1400" HEIGH
                                </TextLine>
```

Table of Contents

2 Situation

- HTR : définition
- Situation
- Qu'est-ce que HTR-United?
- Enjeux et résultats
- Conclusion

Situation de l'HTR et de ses données

- L'OCR et le HTR sont de grandes opportunités pour accéder à des collections de documents et créer des corpus textuels.
- Mais les modèles de transcription sont coûteux à produire car ils nécessitent des données d'entraînement.
- Nous devons nous appuyer sur des modèles et des données préexistants.
- Ils sont rarement FAIR.
  - Difficiles à **trouver** et pas toujours **accessibles**.
  - Formats incertains et annotations variables.
  - Conditions de **réutilisation** peu claires.

Table of Contents

3 Qu'est-ce que HTR-United?

- HTR : définition
- Situation
- Qu'est-ce que HTR-United?
- Enjeux et résultats
- Conclusion

C'est un catalogue

Quelques-uns de nos principes directeurs :

- Consultable par les humains grâce à une interface utilisateur offrant des filtres
- Exploitable par les machines grâce à un catalogue structuré, documenté et versionné synchronisé avec Zenodo
- Un environnement technique minimal pour garantir une maintenance facile

C'est un catalogue ouvert aux soumissions

Le catalogue est alimenté par les créateurs des jeux de données. Comment contribuent-ils ?

- Publication des données (nous offrons un modèle et des lignes directrices pour les bonnes pratiques)
- Création de la nouvelle entrée dans le catalogue (‘htr-united.yml‘) en utilisant notre formulaire
- Interaction via les problèmes Github pour corriger les problèmes que nous repérons lors de la validation de l'entrée

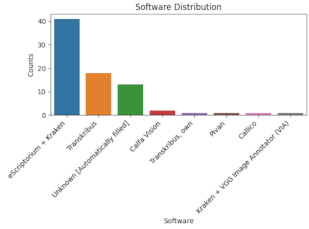
C'est un grand catalogue

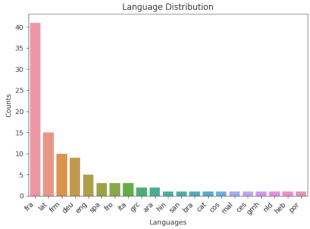
Au 27 mars 2024, le catalogue contient:

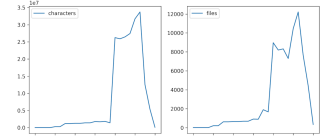
- 81 ensembles de données créés par au moins 37 projets différents
- 21 langues (beaucoup de français et de latin) pour 8 écritures différentes (principalement latin)
- des documents manuscrits et imprimés, mélangés ou "purs"
- une période allant de 800 à 2023
- créés avec au moins 6 logiciels HTR différents
- plus de 43 millions de caractères, plus d'1 million de lignes ou plus de 20 000 images

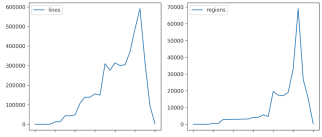
(Résumé graphique de juillet)

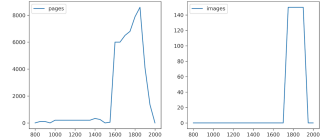
3 Qu'est-ce que HTR-United?

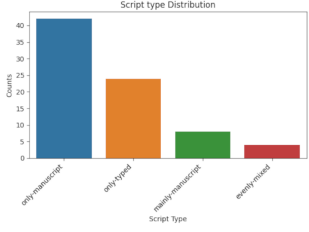


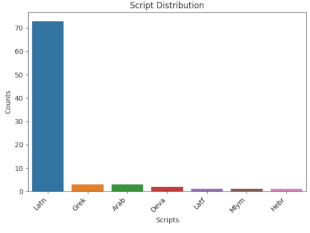












C'est un schéma de description

- Vocabulaire contrôlé
- Formalisé en utilisant JSONSchema
- Contient des métadonnées telles que :
  - Description de la vérité terrain (langue et script, nombre de mains, période couverte, jeu de caractères)
  - Description de l'ensemble de données (lien, titre, description, format de fichier, métriques, licence)
  - Description des conditions de production (projet, auteurs et annotateurs, logiciel)
- Objectif à plus long terme : élaborer un vocabulaire contrôlé pour les lignes directrices de transcription
- Son évolution est transparente et documentée via les problèmes GitHub

C'est un schéma de description ouvert]

3 Qu'est-ce que HTR-United?

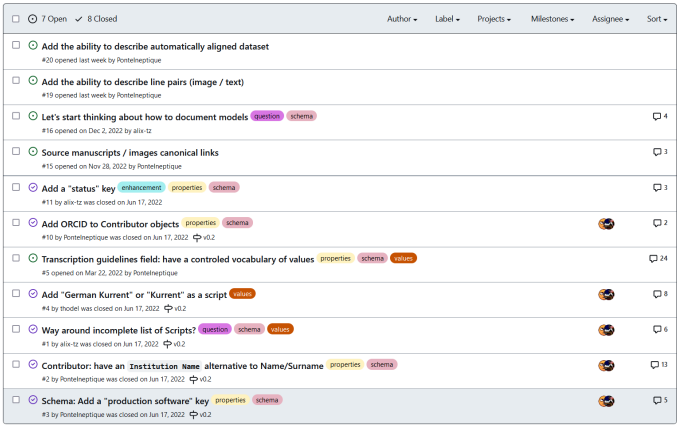


Figure: Sélection d'issues ouvertes sur le dépôt du projet

C'est une boîte à outils

Plusieurs problèmes sont communément rencontrés dans les projets d'HTR / OCR pour la publication des données, donc nous avons créé des outils pour gérer ces derniers plus facilement:

- **HTRuc** : contrôle la validité des fiches de catalogage et permet la compilation du catalogue
- **HTRVX** : contrôle la validité des fichiers XML (y compris les ontologies comme SegmOnto et la présence d'éléments vides)
- **HumGenerator** : calcule des métriques (fichiers, zones, lignes, caractères), crée de jolis badges pour les afficher, met à jour les fiches de catalogage
- **ChocoMufin** : contrôle les caractères dans un ensemble de données, convertit les caractères selon une table de conversion *(initialement développé pour le corpus médiéval CREMMA par A. Pinche et T. Clérice)*

Table of Contents

4 Enjeux et résultats

- HTR : définition
- Situation
- Qu'est-ce que HTR-United?
- Enjeux et résultats
- Conclusion

Enjeux pour la communauté scientifique

- HTR-United contribue à respecter les principes FAIR.
- Il plaide pour la reconnaissance des ensembles de données en tant que résultats scientifiques.
- Il facilite la création de modèles (génériques) sur une plus grande variété de données (car les données sont partagées).
- Il ouvre la voie à une standardisation des pratiques de transcription à travers les plateformes, langues et scripts.
- Il s'est imposé comme catalogue pour les *data management plan* des projets impliquant de l'HTR.

Réutilisations

- HTR-United a déjà été réutilisé "officiellement" (c'est-à-dire cité) par quelques articles en vision assistée par ordinateur, mais il n'est souvent qu'impliqué dans la découverte de jeux de données.
- Quelques "grands" modèles sont issus de son catalogue, en particulier il a permis de connecter des projets dans le cadre de CATMuS (**C**onsistent **A**pproaches to **T**ranscribing **M**an**uS**cripts) et de produire les modèles Print, Modernes et Médiévaux.
- Dans cette lignée, il a permis la publication du dataset CATMuS Médiéval, 160 000 lignes issues de plus de 200 manuscrits pour 10 langues environ sur 800 de productions manuscrites médiévales.

Réutilisations

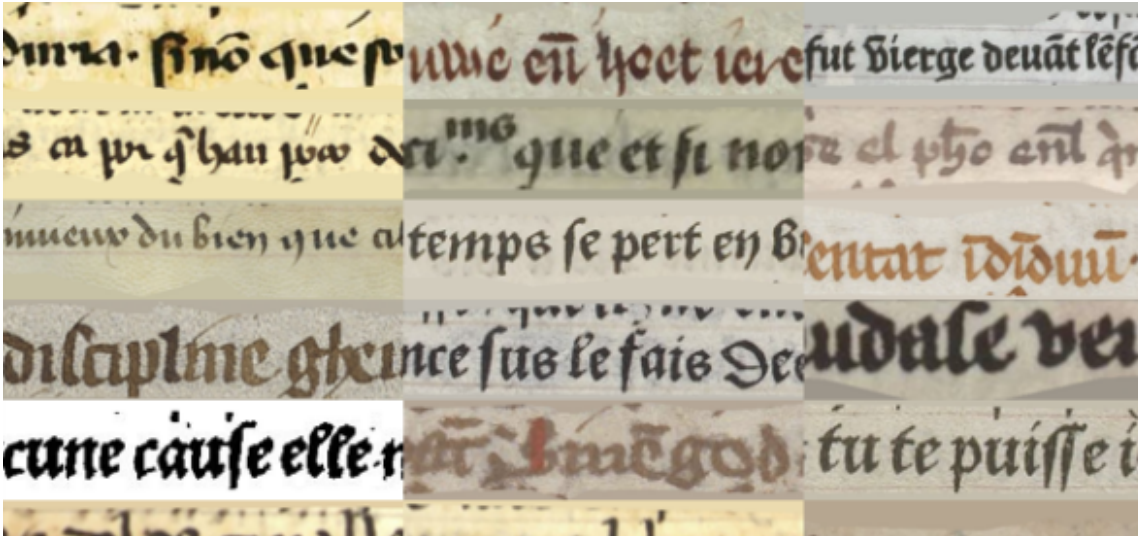


Table of Contents

5 Conclusion

- HTR : définition
- Situation
- Qu'est-ce que HTR-United?
- Enjeux et résultats
- Conclusion

Le futur

5 Conclusion

- Nous souhaitons adosser à HTR-United un système de *data papers* via un partenariat avec un journal.
- Améliorer l'interface du catalogue
- Rendre "plus" *machine actionable* le catalogue pour créer des datasets ad-hoc.