DEEP-AGORA

Analyse et spécifications

TABLE DES MATIÈRES

01

Contexte

Client, enjeux et besoins

CLIENT

CESR : Centre d’Etudes Supérieures de la Renaissance

Propose masters et doctorats en:

- Histoire, Civilisation, Patrimoine.
- Humanités Numériques.

ENJEUX

BVH

Programme de valorisation régionale d'ouvrages anciens

BaTyr

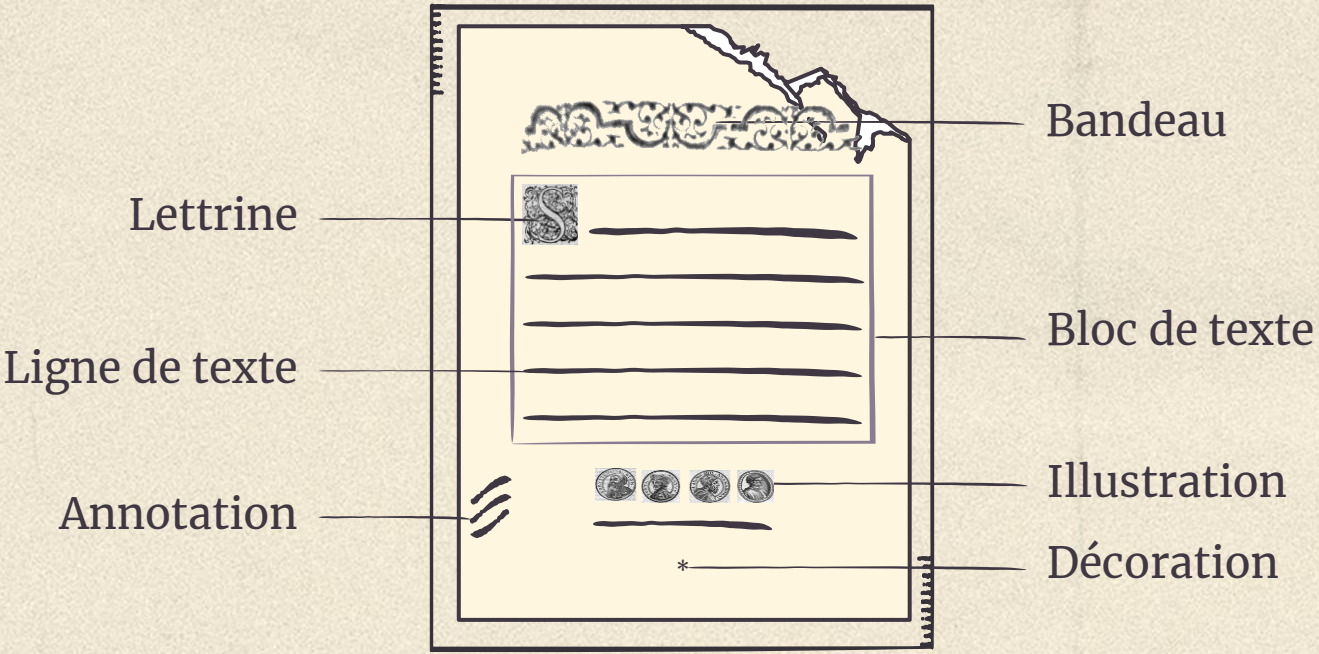
Base de données d'illustrations extraites d'ouvrages

TypoRef

Recherche de spécimens de polices d'écriture

BESOINS

Segmenter un document



BESOINS

Eléments de contenu

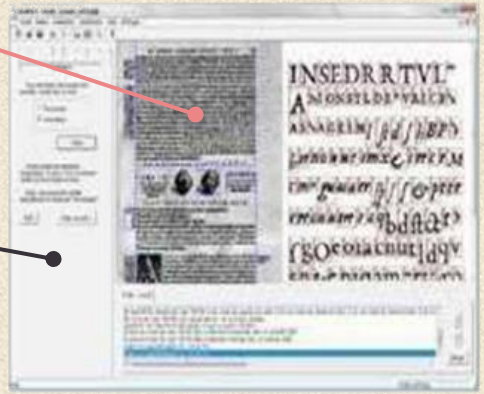
Texte			Texte & image	Image		
Bloc de texte	Ligne	Annotation manuscrite	Lettrine	Bandeau	Illustration	Décoration

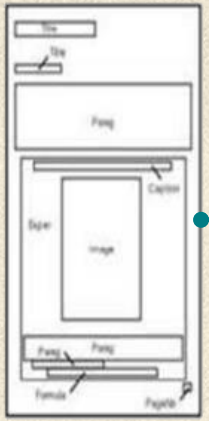
BESOINS

Intéractivité

Document ancien

Scénario



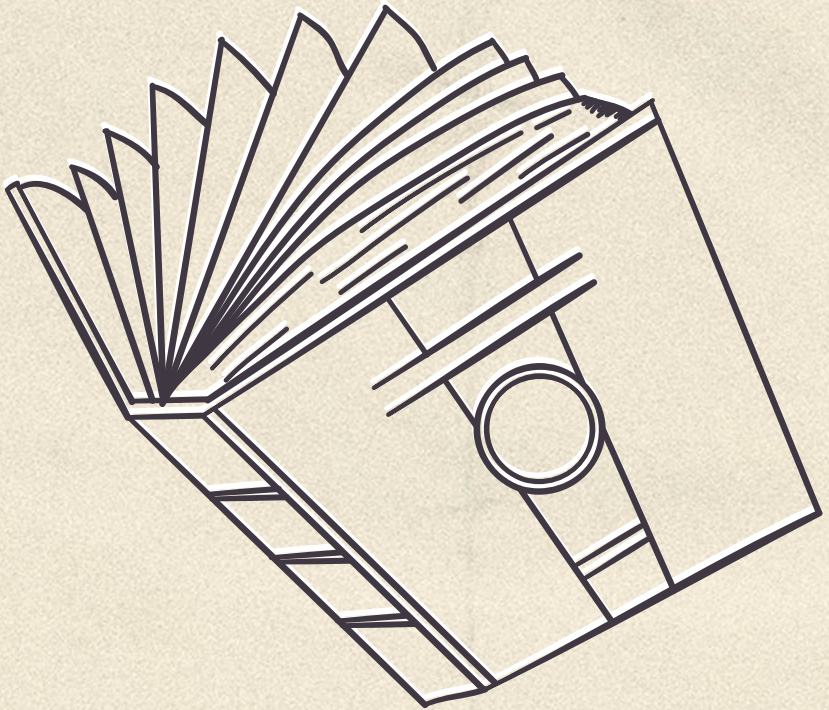


Elements de contenu

02

Existant

Système et limitations



SCENARIOS

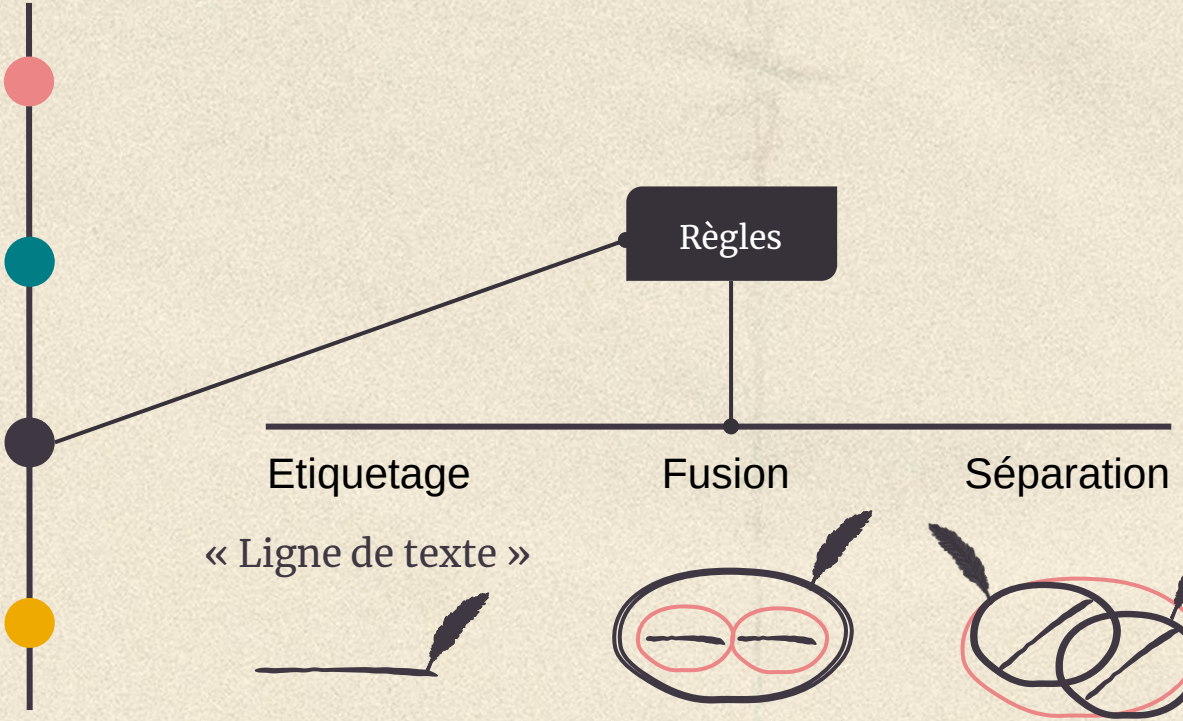
Importation d'image

Binarisation

Passe l'image en N&B

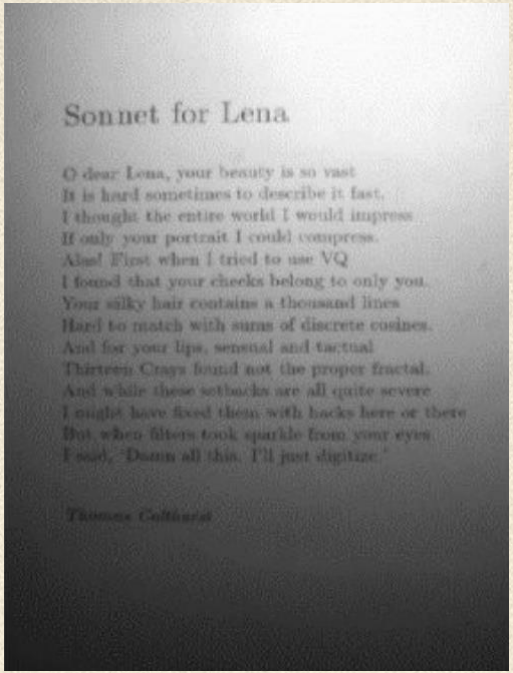
Définition de règles

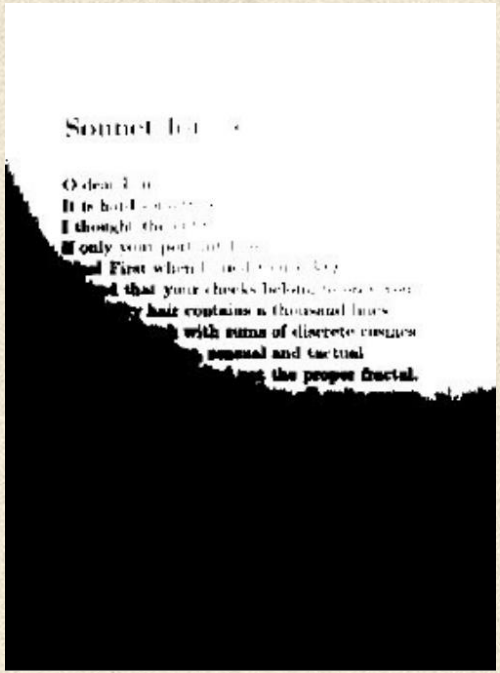
Exportation au format ALTO



LIMITES

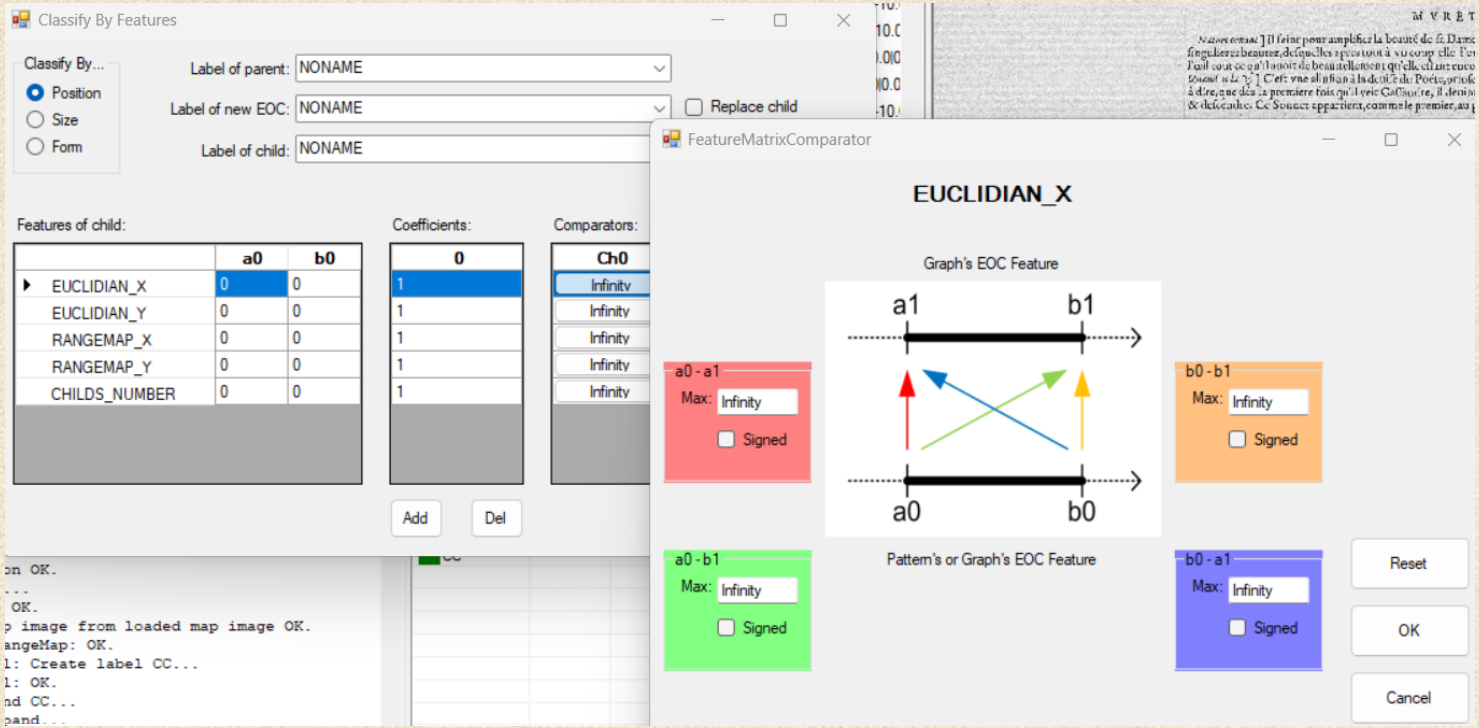
Binariser...





LIMITES

Définir des règles...



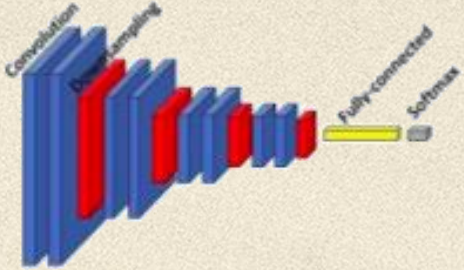
03

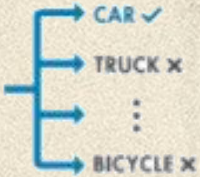
Etat de l’art

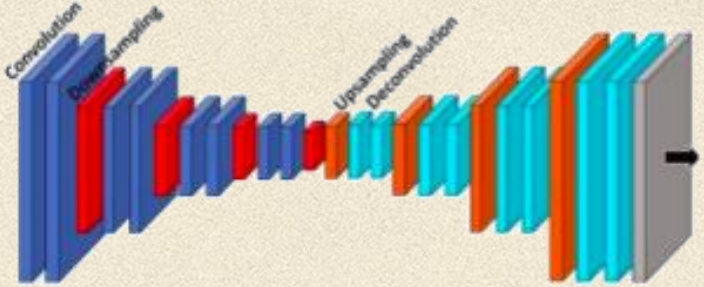
Vision par ordinateur et outils pour documents anciens

VISION PAR ORDINATEUR

Les réseaux de neurones convolutifs

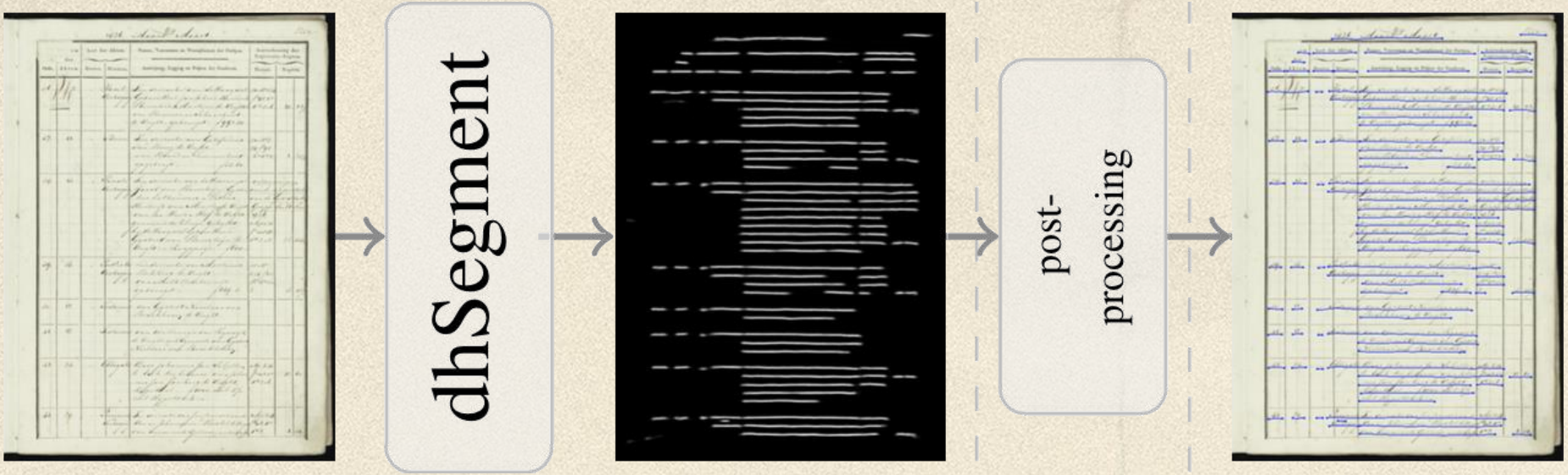






OUTILS POUR DOCUMENTS ANCIENS

Le Framework dhSegment



04

Nouveautés

Résolution des problèmes et refonte

RÉSOLUTION DES PROBLÈMES

Binarisation

Définition de règles

Ecriture manuscrite

Convolutions

L'architecture du réseau de neurones ne l'utilise pas

Optimisation

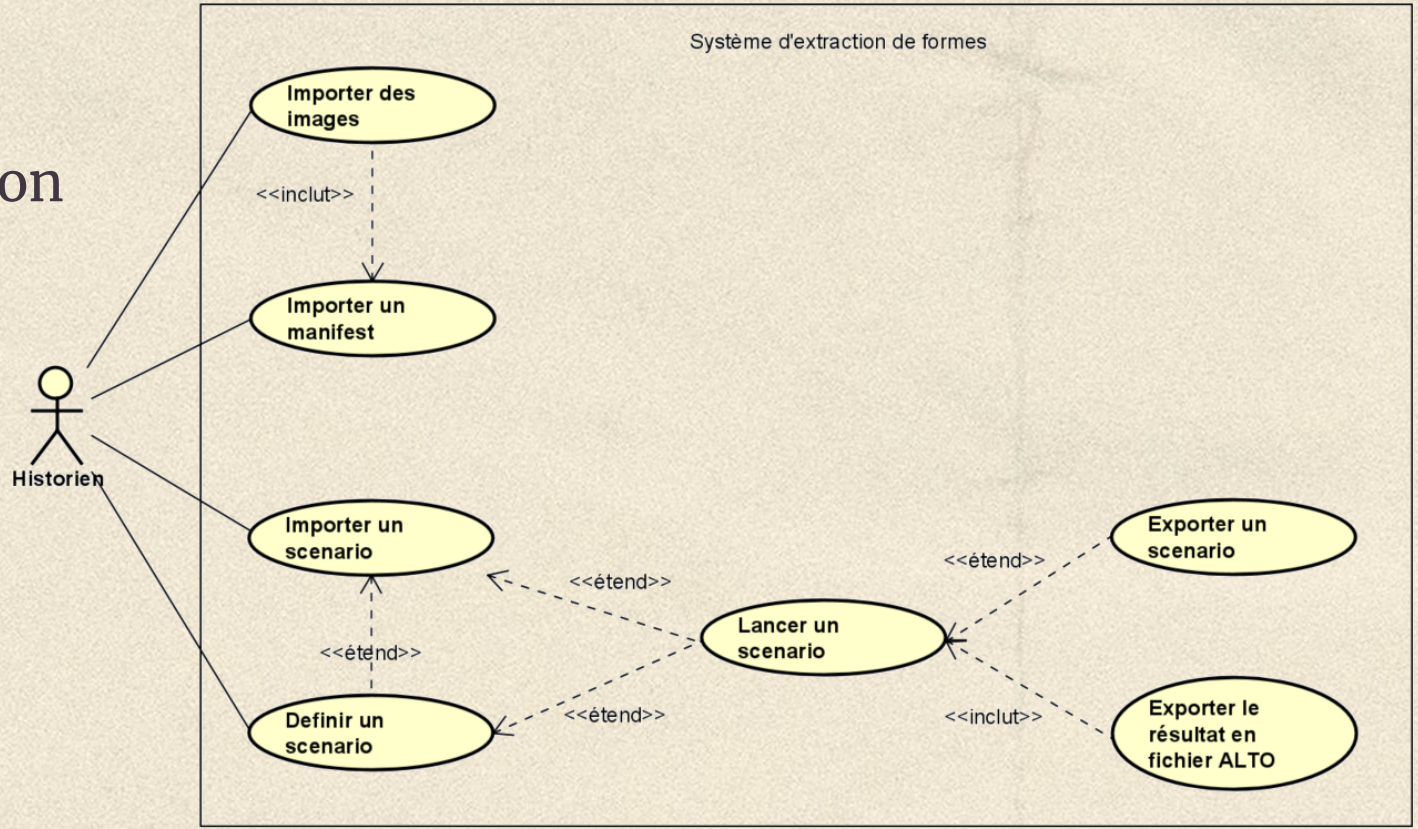
Les meilleures règles sont apprises automatiquement

Jeux de données

L'entraînement se fait aussi sur des documents manuscrits

REFONTE

Maintient des cas d'utilisation



05

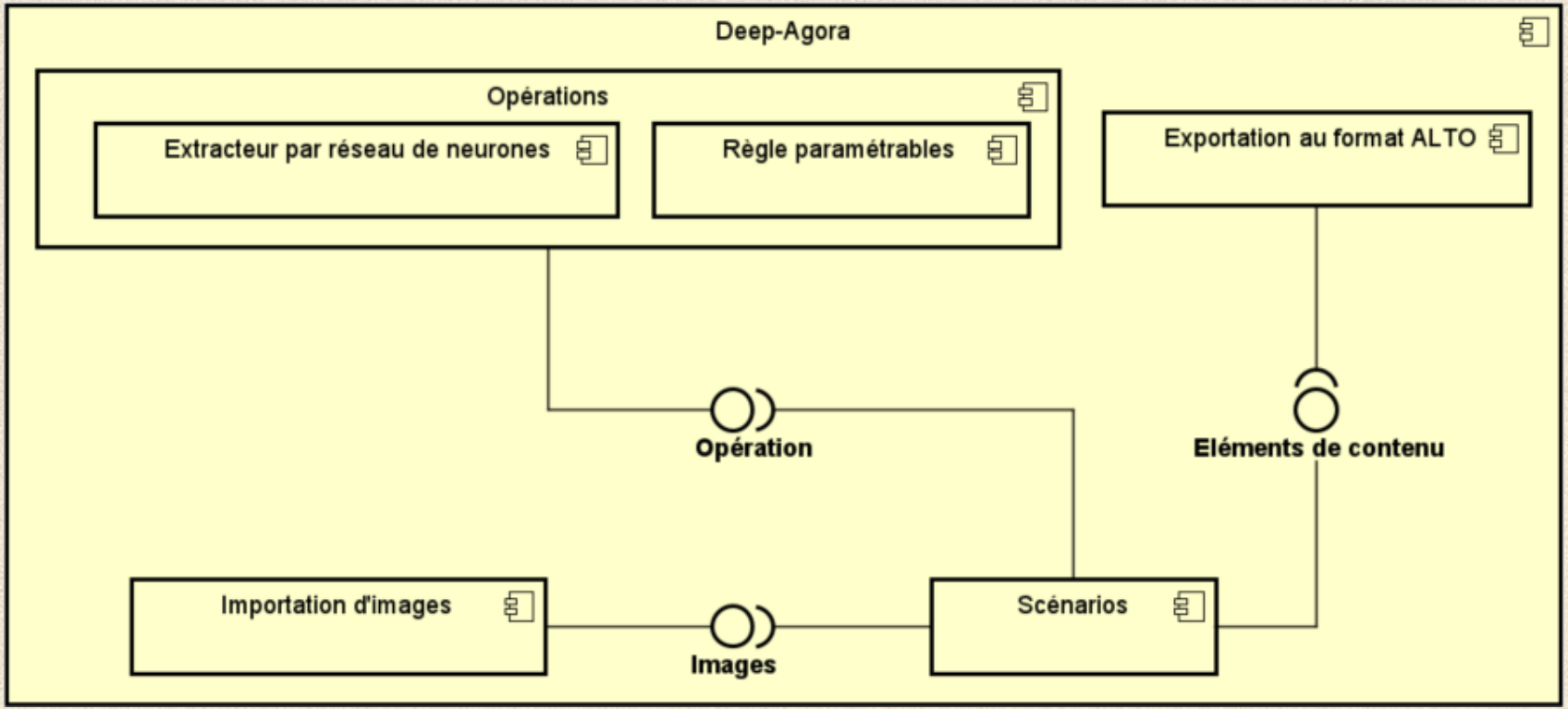
Conception

Composants, méthodologie et données

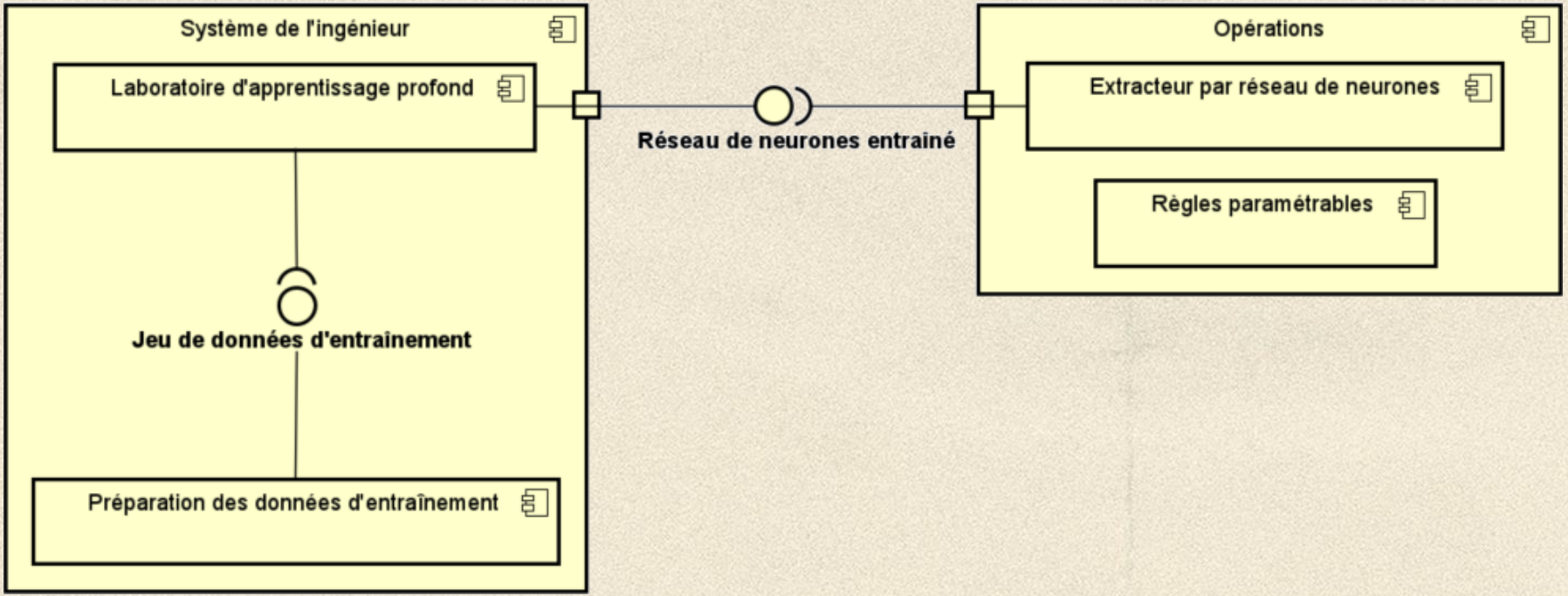
JEUX DE DONNÉES D'APPRENTISSAGE

	Binarisé	Manuscrit/imprimé	Mise en page	Ligne de texte	Illustrations	Nombre de pages	Format
IMPACT		Both	x	x	x	600 K	PAGE XML
ICFHR18 RASM2018		Handwritten	x	x	x	65	PAGE XML
ICDAR19 RASM2019		Handwritten	x	x	x	120	PAGE XML
HORAE		Handwritten	x	x	x	797	PAGE XML
ESPOSALLES		Handwritten	x	x		173	N/A
FCR		Handwritten	x	x		500	PAGE XML
DIVA-HisDB		Handwritten	x	x		150	PAGE XML
Pinkas		Handwritten	x	x		30	PAGE XML

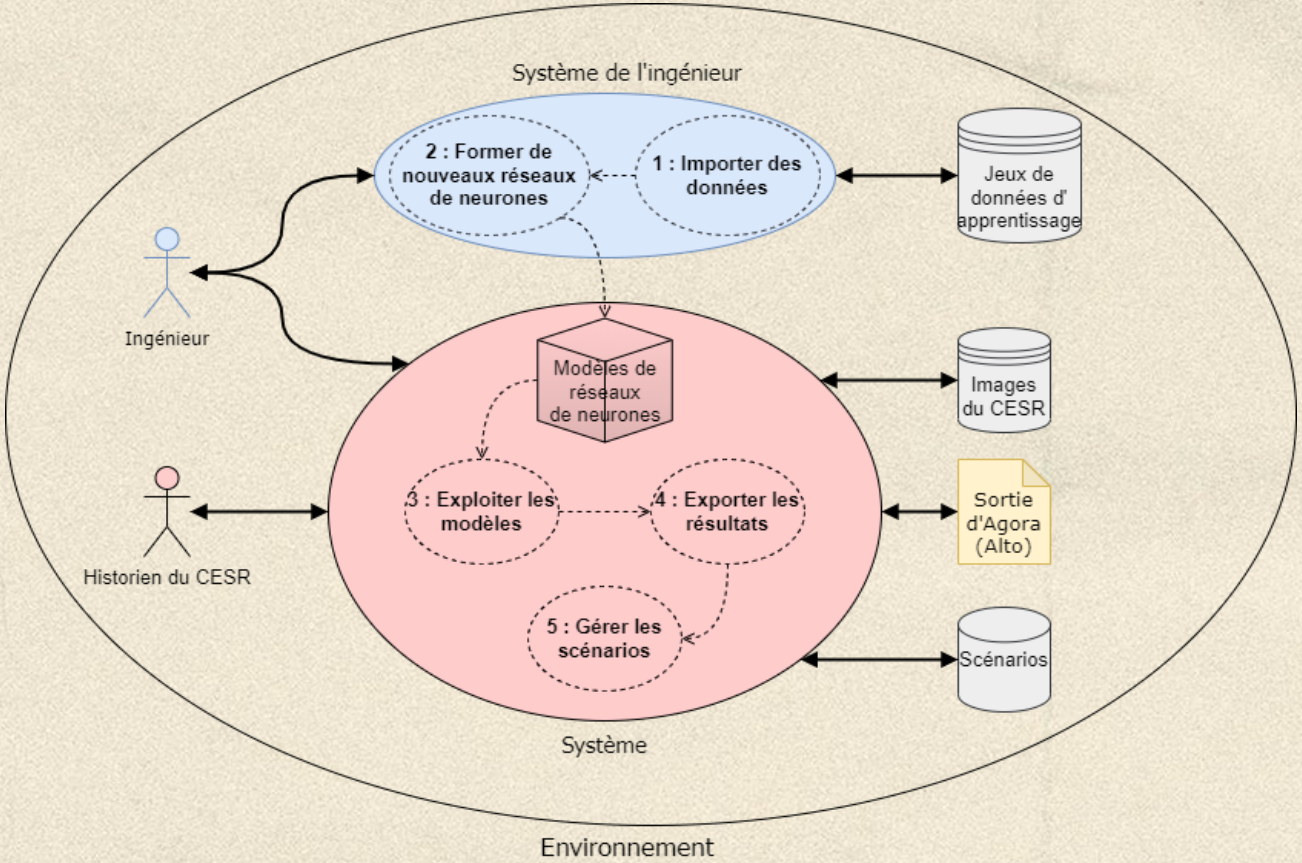
DEEP-AGORA


Deep-Agora
Opérations
Extracteur par réseau de neurones
Règle paramétrables
Exportation au format ALTO
Opération
Eléments de contenu
Importation d'images
Images
Scénarios

CRÉATION DES RÉSEAUX DE NEURONES



MÉTHODOLOGIE



MODULES A DEVELOPPER

Module d'importation des données d'apprentissage

Laboratoire d'apprentissage profond

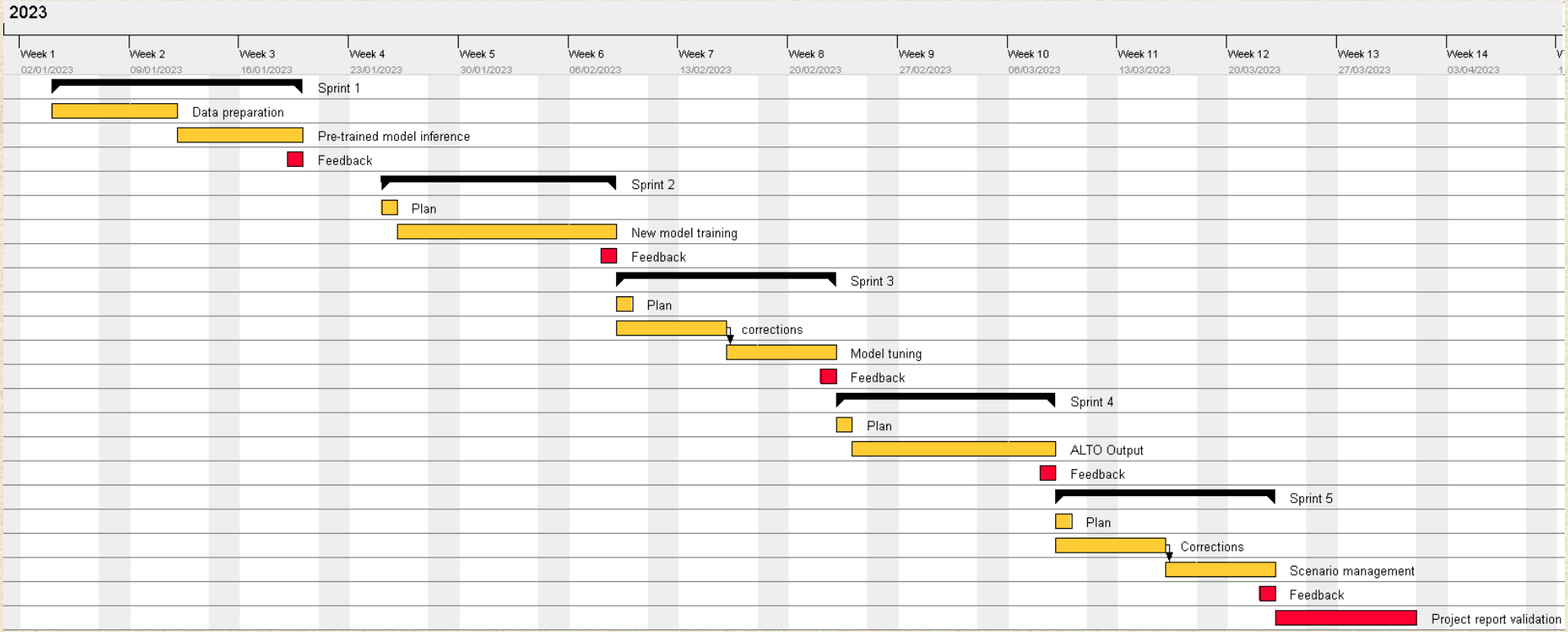
Module d'importation des images du client

Module de mise en forme des prédictions

Module d'exportation des résultats

Module de gestion des scénarios

PLANIFICATION



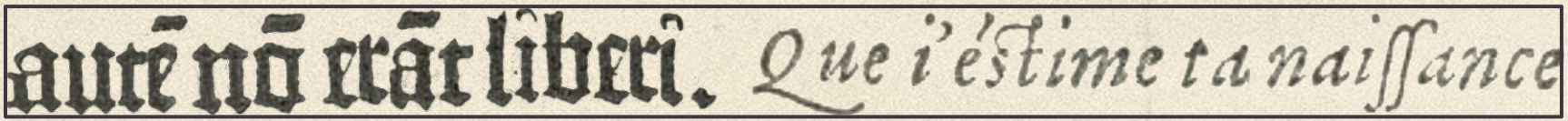
06

Analyse critique

Hypothèses et risques

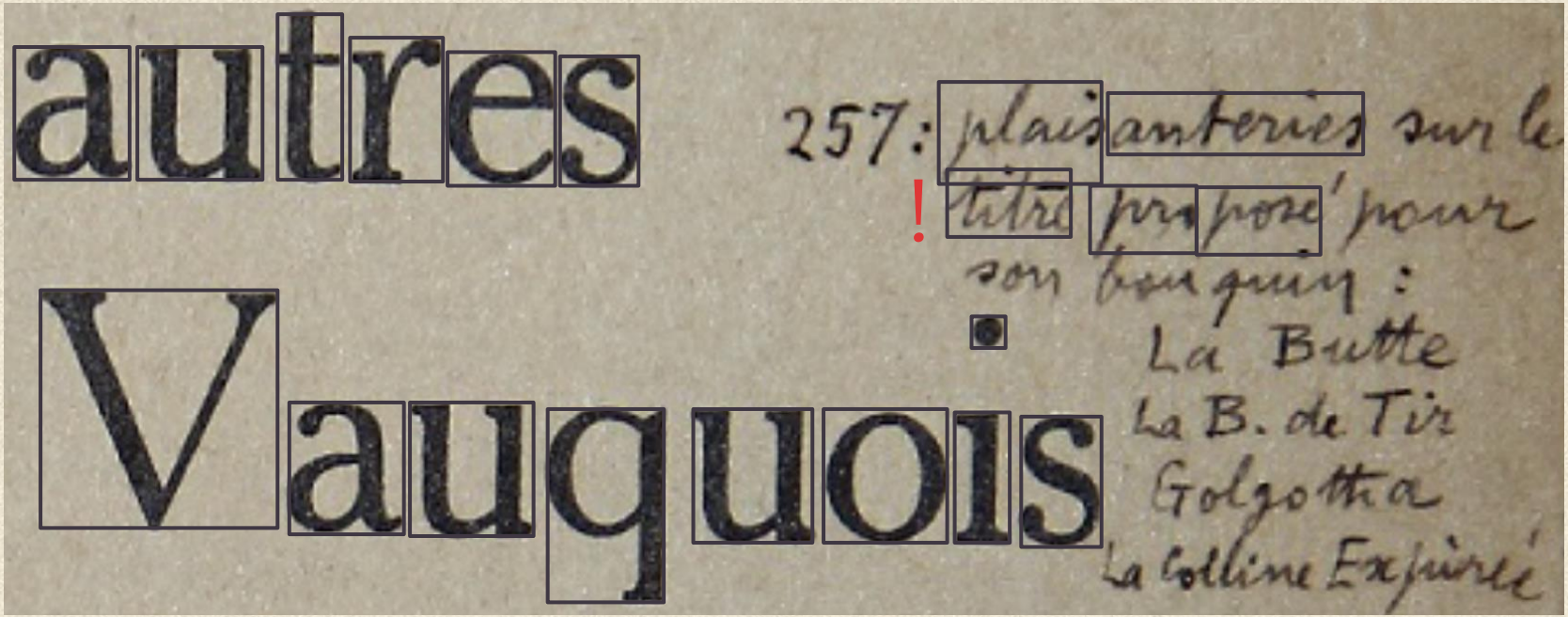
HYPOTHÈSES

Les lignes de texte peuvent être à police hétérogène



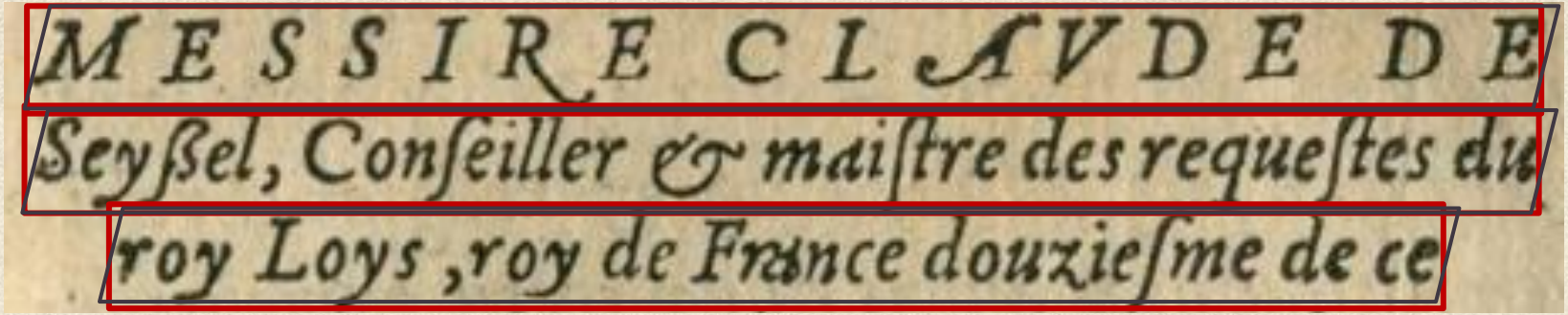
HYPOTHÈSES

Pas de segmentation de caractères en écriture manuscrite



RISQUES

Le format ALTO...




```
<TextLine HEIGHT="60" HPOS="108" ID="PAG_00000013_TL000004" STYLEREFS="TXT_2" VPOS="328" WIDTH="1208"
...
</TextLine>
<TextLine HEIGHT="58" HPOS="108" ID="PAG_00000013_TL000005" STYLEREFS="TXT_2" VPOS="394" WIDTH="1208"
...
</TextLine>
<TextLine HEIGHT="62" HPOS="110" ID="PAG_00000013_TL000006" STYLEREFS="TXT_2" VPOS="461" WIDTH="1209"
...
</TextLine>
```

RISQUES

Pas assez de données ?

Volume

L'entraînement se fait sur un grand volume de données

Diversité

De nombreux cas différents pour éviter les biais

	Nombre de pages
IMPACT	600 K
ICFHR18 RASM2018	65
ICDAR19 RASM2019	120
HORAE	797
ESPOSALLES	173
FCR	500
DIVA-HisDB	150
Pinkas	30

RISQUES

Pas assez de données ?

39

Jeux de données différents au total

IMPACT [129]
ICFHR18 RASM2018 [34]
ICDAR19 RASM2019
HORAE [15]
GERMANA [141]
RODRIGO [159]
ESPOSALLES [ ]
BH2M [51]
FCR [143]
HisClima [150]
DIVA-HisDB [168]
Pinkas [93]
ICDAR19 HDRC-Chinese [155]
ENP [32]
ICDAR17 REID2017 [33]
ICDAR19 DMAS201955
ICDAR19 REID2019 [35]
Newspaper Navigator [99]
HBA 1.0 [113]
HADARA80P [128]
Tripitaka [189]
Oficio de Hipotecas de Girona (OHG)
ABP & NAF [133]
HJDataset [166]
BIR-database [8]
IlluHisDoc [118]
PHTD [2]
PBOK [5]
GRPOLY-DB [62]
SleukRith [181]
Lontar Sunda [172]
ICFHR18 Asian Palm Leaf [87]
HTR Benchmarks [177]
Digital Peter [110]
BiblIA [42]
POPP [39]
READ-BAD [68]
Warped Arabic [47]
BADAM [88]

RISQUES

Infrastructure

> 32,8M paramètres au total [...]
> 9,36M doivent être entièrement entraînés.
>
> — ***dhSegment: A generic deep-learning approach for document segmentation***

Merci

Avez-vous des questions ?

theo.boisseau@etu.univ-tours.fr