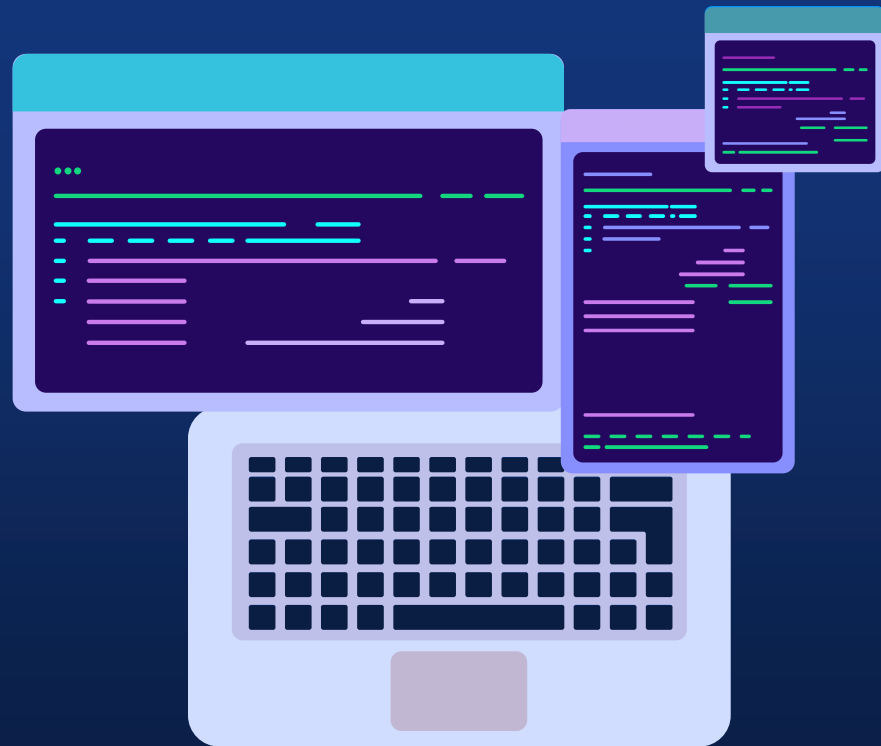


# DEEP-AGORA

Existant &  
État de l'art



# TABLE DES MATIÈRES

01

EXPRESSION DES BESOINS

02

EXISTANT

03

À FAIRE ÉVOLUER

04

ETUDE DE FAISABILITÉ

05

ANALYSE DES LIBRAIRIES

06

À VENIR





01

EXPRESSION  
DES BESOINS



# UTILISATEUR FINAL



Centre d'Etudes Supérieures de la Renaissance

## Le CESR

Centre de formation et de recherche  
situé à Tours.

Propose masters et doctorats en:

- Histoire, Civilisation, Patrimoine.
- Humanités Numériques.



# QUELQUES EXEMPLES D'EXTRACTIONS



...aenean e  
s intègèr a  
ra socios



02

EXISTANT





# AGORA (actuellement)

Présenter le  
standard ALTO

## BINARISATION

Choix entre différents  
algorithmes

## FICHIERS ALTO (XML)

Extraction des éléments  
de contenus (EOC) avec  
leurs étiquettes dans un  
système de fichier

## SCÉNARIOS


Destinés à regrouper  
et étiqueter les pixels  
noirs à partir de règles

## 2 CAS D'UTILISATION

Stockés en base de  
données ou passés à  
RETRO (OCR)

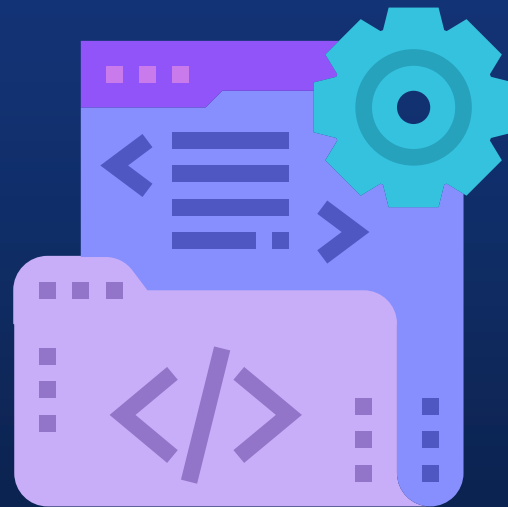






# 03

À FAIRE  
ÉVOLUER



# PROBLÈMES VS SOLUTION



## BINARISATION + ÉCRITURE MANUSCRITE + DIFFICILE À MAITRISER

- Binarisation pas suffisamment efficace
- Des caractères et des lignes qui se touchent
- Des interfaces trop complexes



## MODULE DE DEEP-LEARNING

- Remplace la binarisation
- Adaptabilité à l'écriture manuscrite
- Adaptabilité à davantage de corpus

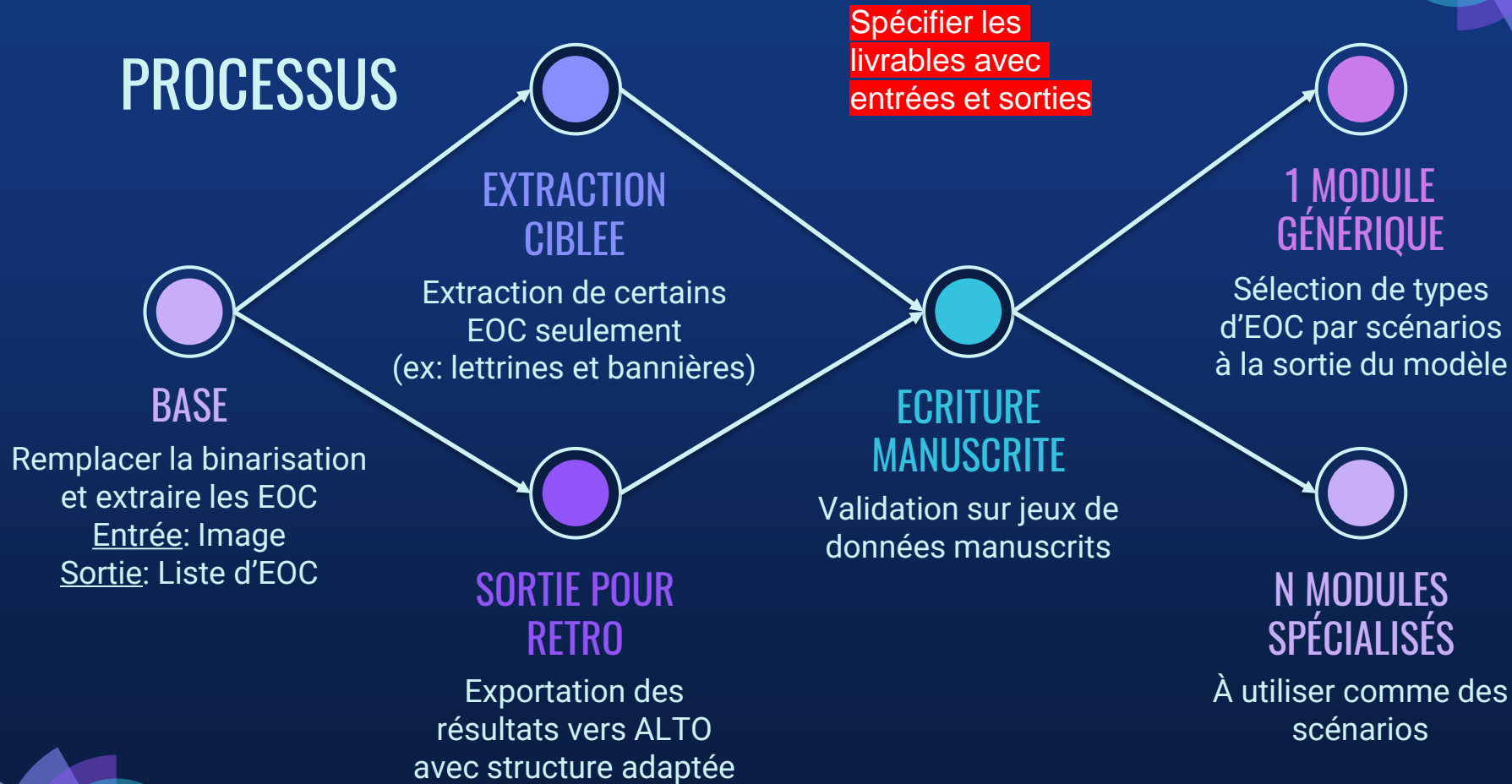


# 04

## ETUDE DE FAISABILITÉ



# PROCESSUS





## 1 MODULE GÉNÉRIQUE

Sélection de types  
d'EOC par scénarios  
en sortie du modèle



## N MODULES SPÉCIALISÉS

À utiliser comme des  
scénarios



Les utilisateurs  
créent leurs  
propres  
modèles

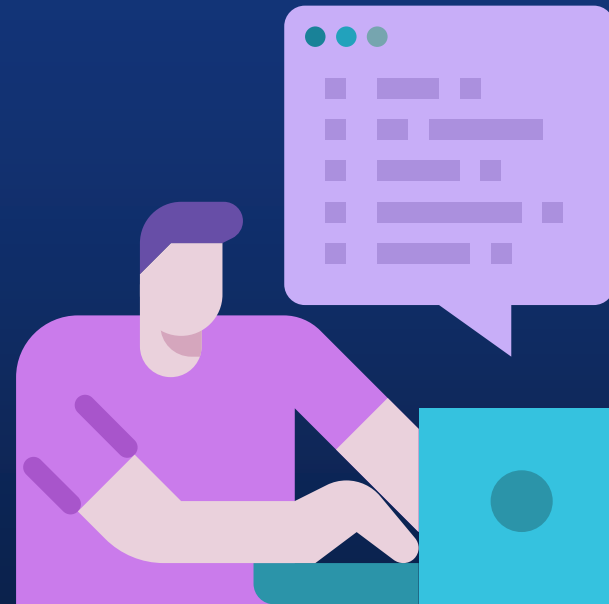


Possibilité  
d'intégrer des  
scénarios au  
modules



# 05

## ANALYSE DES LIBRAIRIES



# Trouver des librairie adaptées

## ANCIENT DOCUMENT LAYOUT ANALYSIS

Conçue pour les documents  
anciens spécifiquement.



## OPEN SOURCE

Code source accessible  
pour pouvoir être modifié.

## TRANSFER LEARNING

Un modèle pré-entraîné  
serait un avantage.



## ADAPTABILITÉ

Documentation solide et  
bonnes pratiques.





# Les plus adaptées

Layout Parser	Kraken	dhSegment
Conçu pour les documents imprimés	Binarise les images et segmente des lignes de texte	Documentation limitée et pas de sortie ALTO
Communauté active et excellente documentation	Documents imprimé / manuscrits, dans divers langages et sortie en ALTO	Documents imprimés / manuscrits et extrait les images



# Choix d'une librairie

Spécifier les  
colonnes, les  
détailler et intégrer  
les types d'EOC



	DOCUMENTS MANUSCRITS	SANS ALGO DE BINARISATION	SORTIE XML ALTO	EXTRACTION D'IMAGES
LAYOUT PARSER				
KRAKEN				
DHSEGMENT				





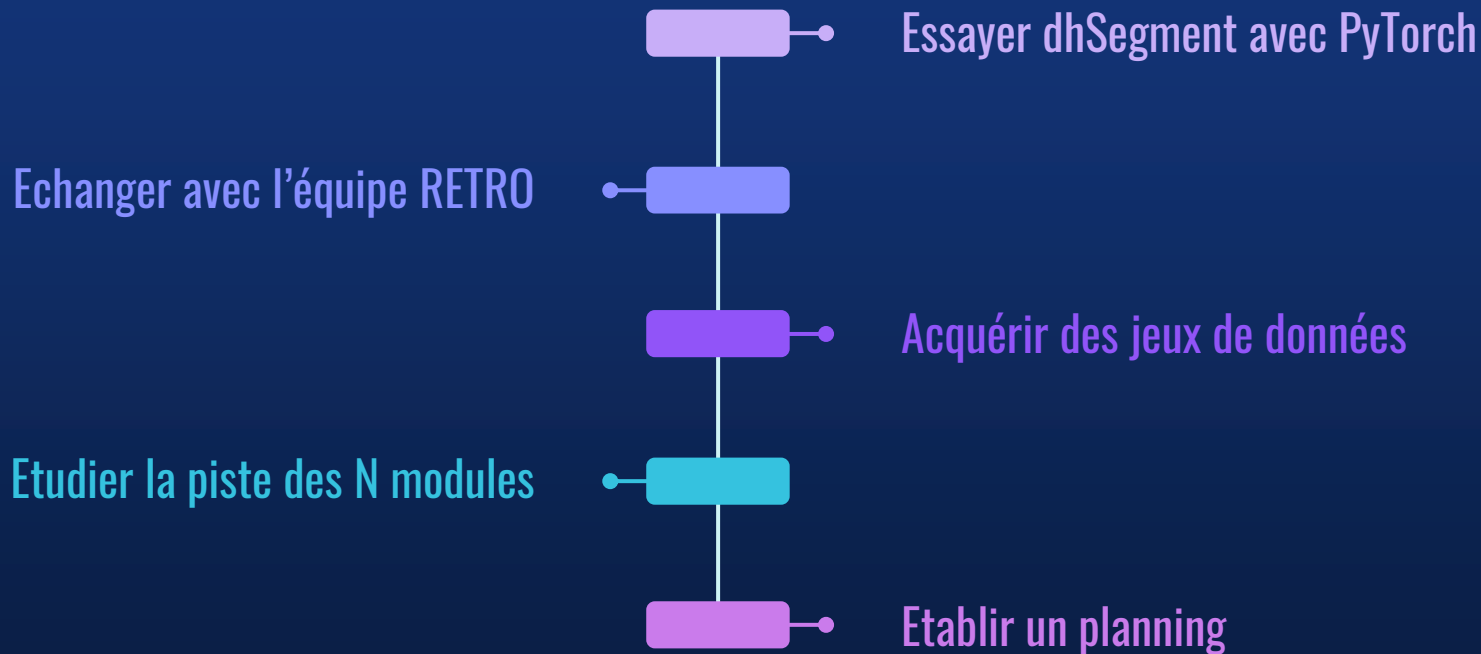
# 06

À VENIR



# À RÉALISER TRÈS PROCHAINEMENT

Spécifier les livrables  
plutôt



# MERCI



Avez-vous des questions ?

[theo.boisseau@etu.univ-tours.fr](mailto:theo.boisseau@etu.univ-tours.fr)  
Polytech Tours

**Please keep this slide for attribution**

**CREDITS:** This presentation template was created by **Slidesgo**, including icons by **Flaticon**, infographics & images by **Freepik**

