

# COMMENT INTRODUIRE SIMPLEMENT ET UNIFORMÉMENT DEUX MODES D'INTERACTION ASYNCHRONES COMPLÉMENTAIRES DANS UN SYSTÈME D'ANALYSE DE DOCUMENTS EXISTANT

Joseph CHAZALON, Bertrand COÜASNON, AURÉLIE LEMAITRE

[www.irisa.fr/intuidoc](http://www.irisa.fr/intuidoc)



**Yvelines**  
Conseil général



ARCHIVES  
YVELINES

*Financement*



UNIVERSITÉ DE  
**RENNES 1**



UNIVERSITÉ  
**RENNES**

*Employeurs*



UMR

**IRISA**

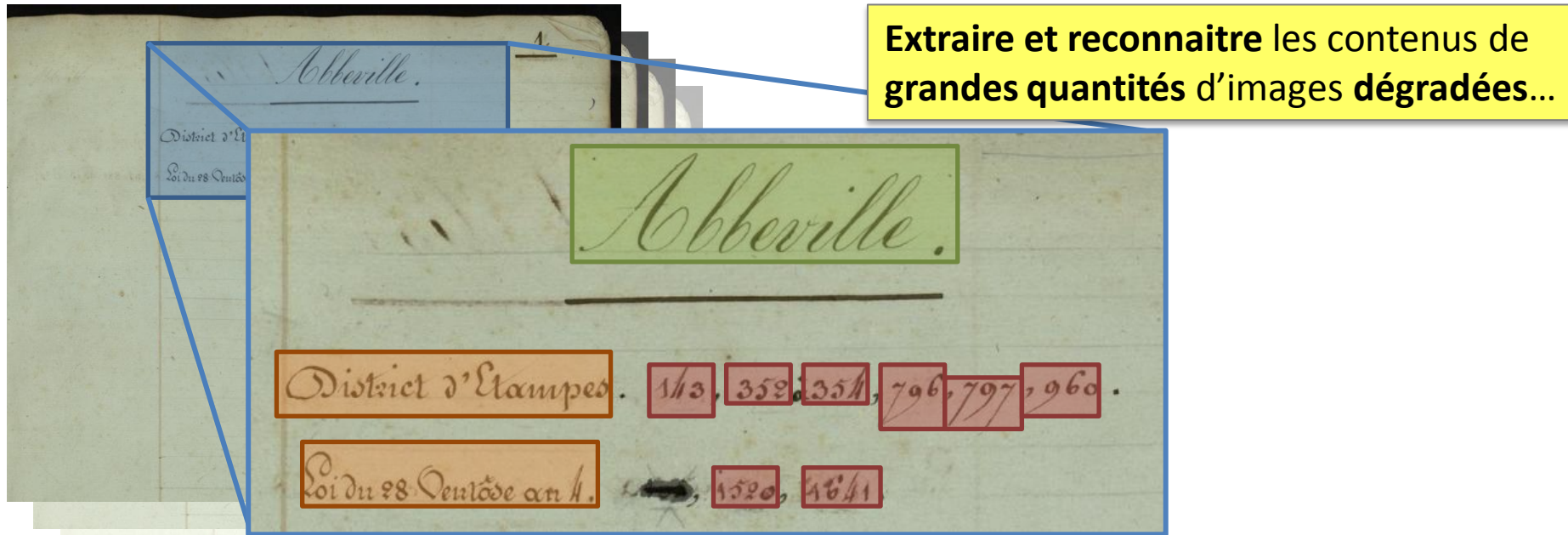
*Unité de recherche*



UNIVERSITÉ  
EUROPÉENNE  
DE BRETAGNE

*PRES*

# Contexte: analyse de documents anciens

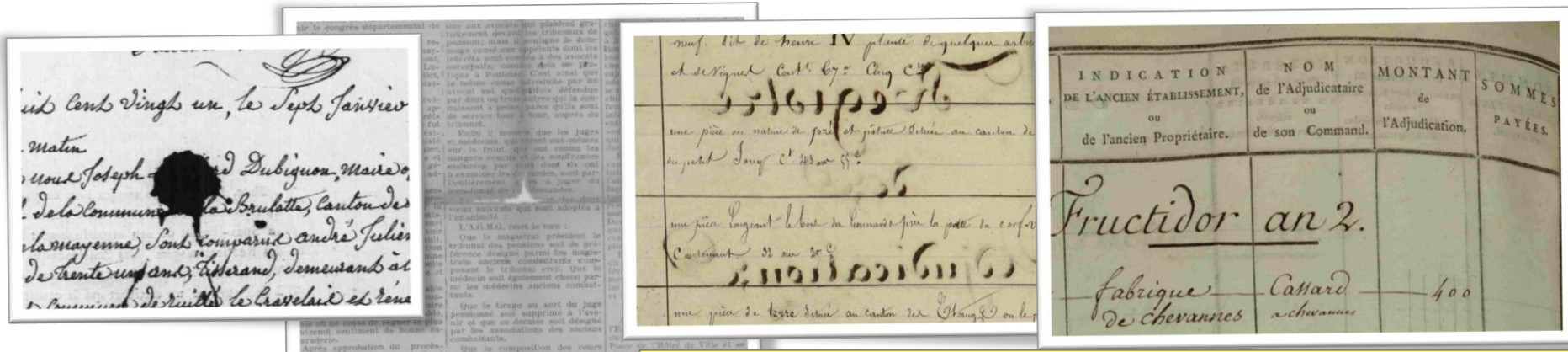


... pour les **indexer** en base de données

Image file	Town / Place	Kind of sale	Sale reference
00071.jpg	Abbeville	dist_etampes	143
00071.jpg	Abbeville	loi_ventose	1520

# Difficultés variées → Erreurs nombreuses

## Dégradations



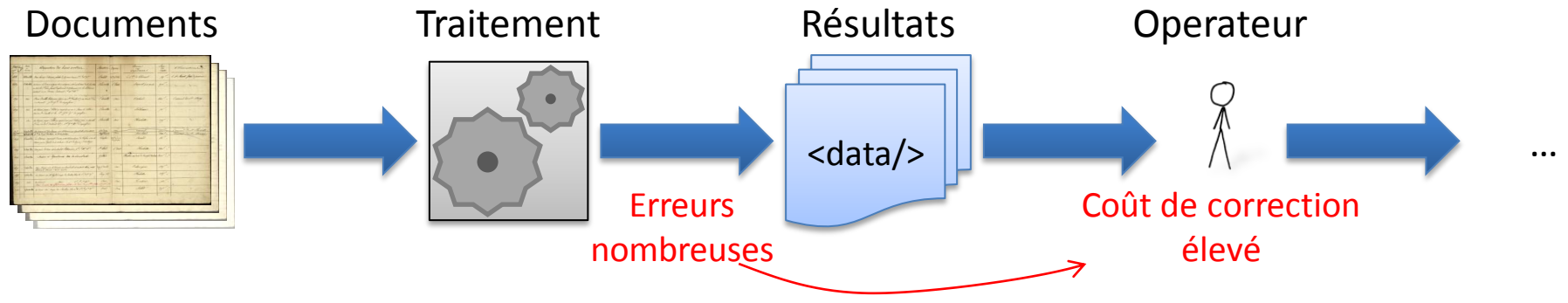
→ Résultats imprévisibles  
→ Intervention humaine nécessaire

## Variabilités

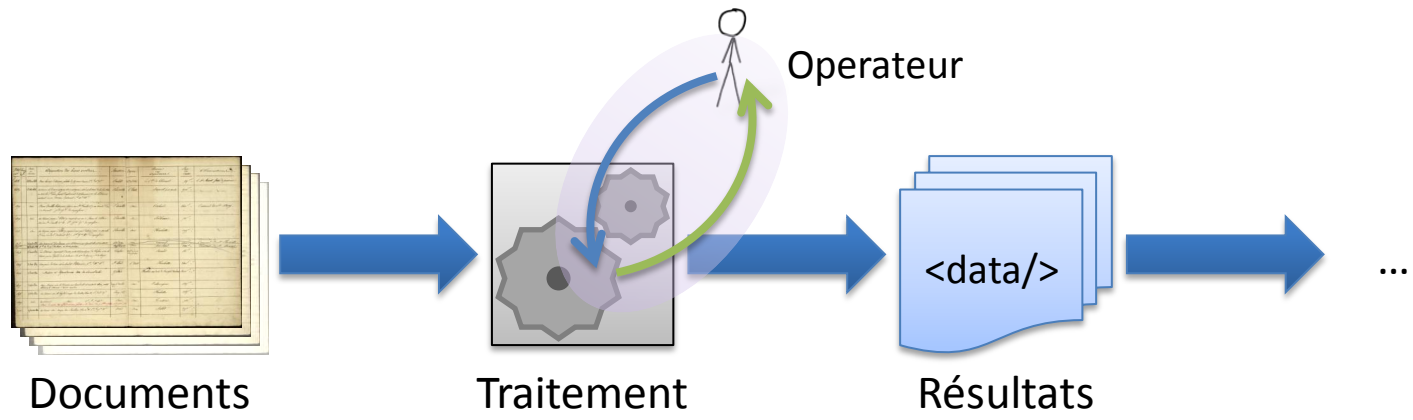


# Pourquoi interagir pendant la phase d'analyse ?

Développer – traiter – corriger : **inefficace** pour documents dégradés



Interagir **pendant** l'analyse pour **diminuer le coût de correction**



- ✓ Correction au plus tôt (propagation limitée)
- ✓ Amélioration (résultats passés et futurs)

# Notre travail

## Objectifs

- Utiliser l'information **externe** pour **améliorer** les réponses
- Autoriser une interaction **asynchrone**
  - Humain / système ne doivent pas s'attendre mutuellement
- Permettre une « conception synchrone »
  - Perturber le moins possible les descriptions de documents

## Contributions

- Définition d'un framework d'**analyse asynchrone**
  - Adapté au traitement de documents anciens en grand volume
- Gestion des erreurs grâce à **2 modes d'interaction**
  - Selon que la détection d'erreur automatique soit possible ou non
- Recommandations pour **adapter un analyseur de pages** existant
  - Pauses/reprises de l'analyse gérées avec mécanisme standard d'exceptions

# Contenu de cette présentation

## 1. Comment rendre un analyseur de pages interactif ?

Transformation en 3 volets

## 2. Exemple 1 : Interaction dirigée

Transcription de patronymes manuscrits

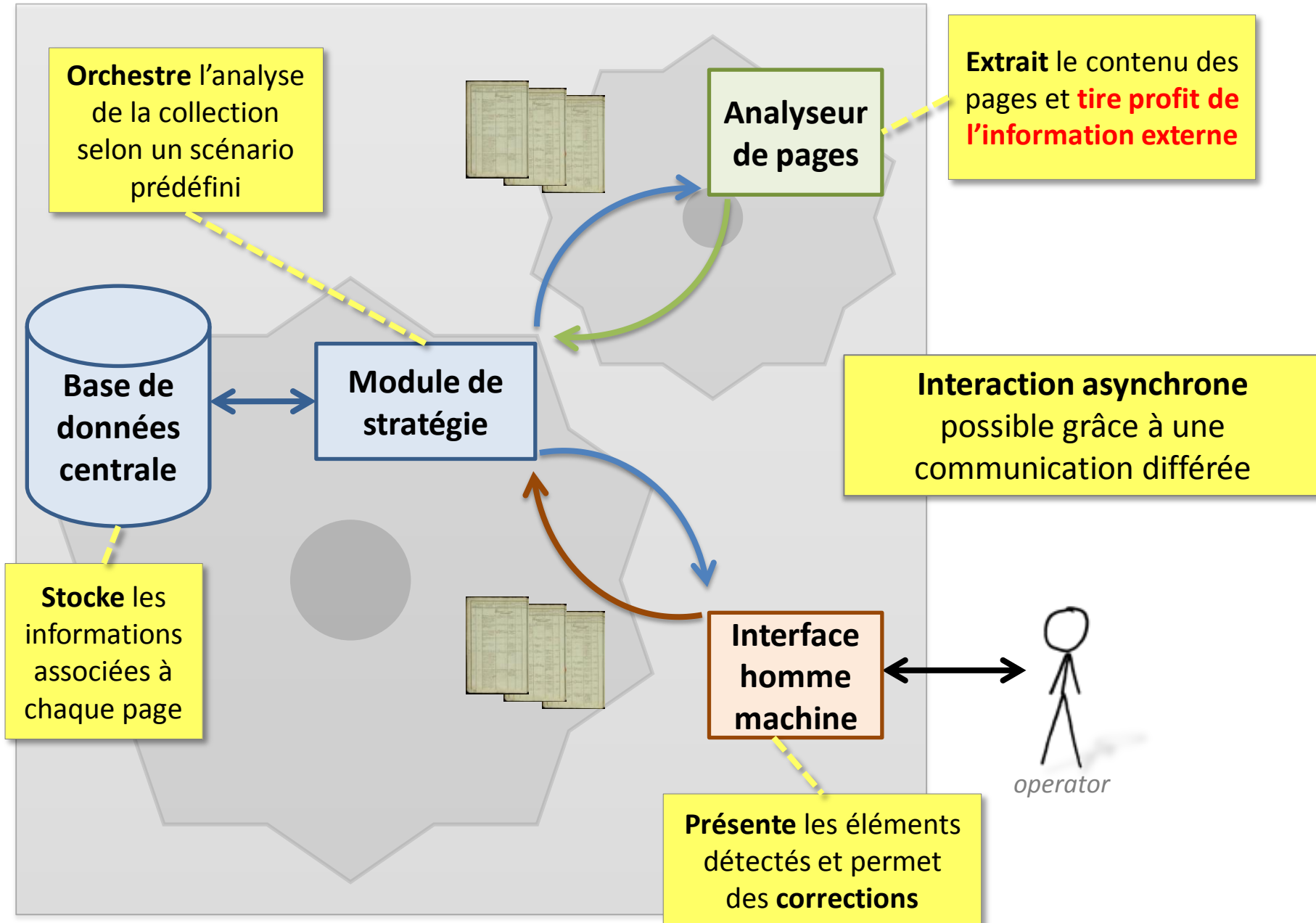
## 3. Exemple 2 : Interaction spontanée

Correction de sous-segmentation de nombres manuscrits

Rendre possible une interaction asynchrone

**COMMENT RENDRE UN ANALYSEUR  
DE PAGES INTERACTIF ?**

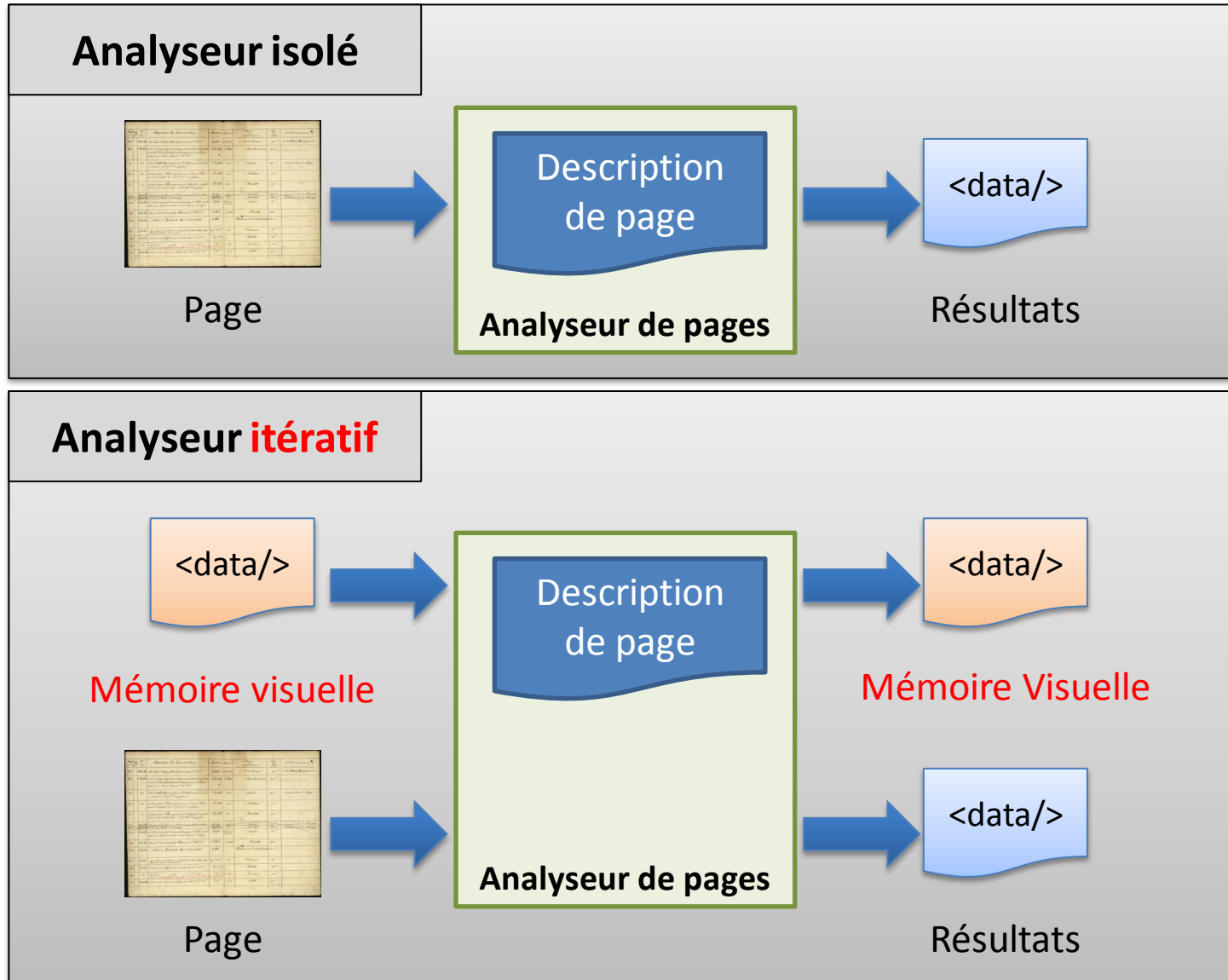
# 1/ Intégration dans l'analyse de « collection »





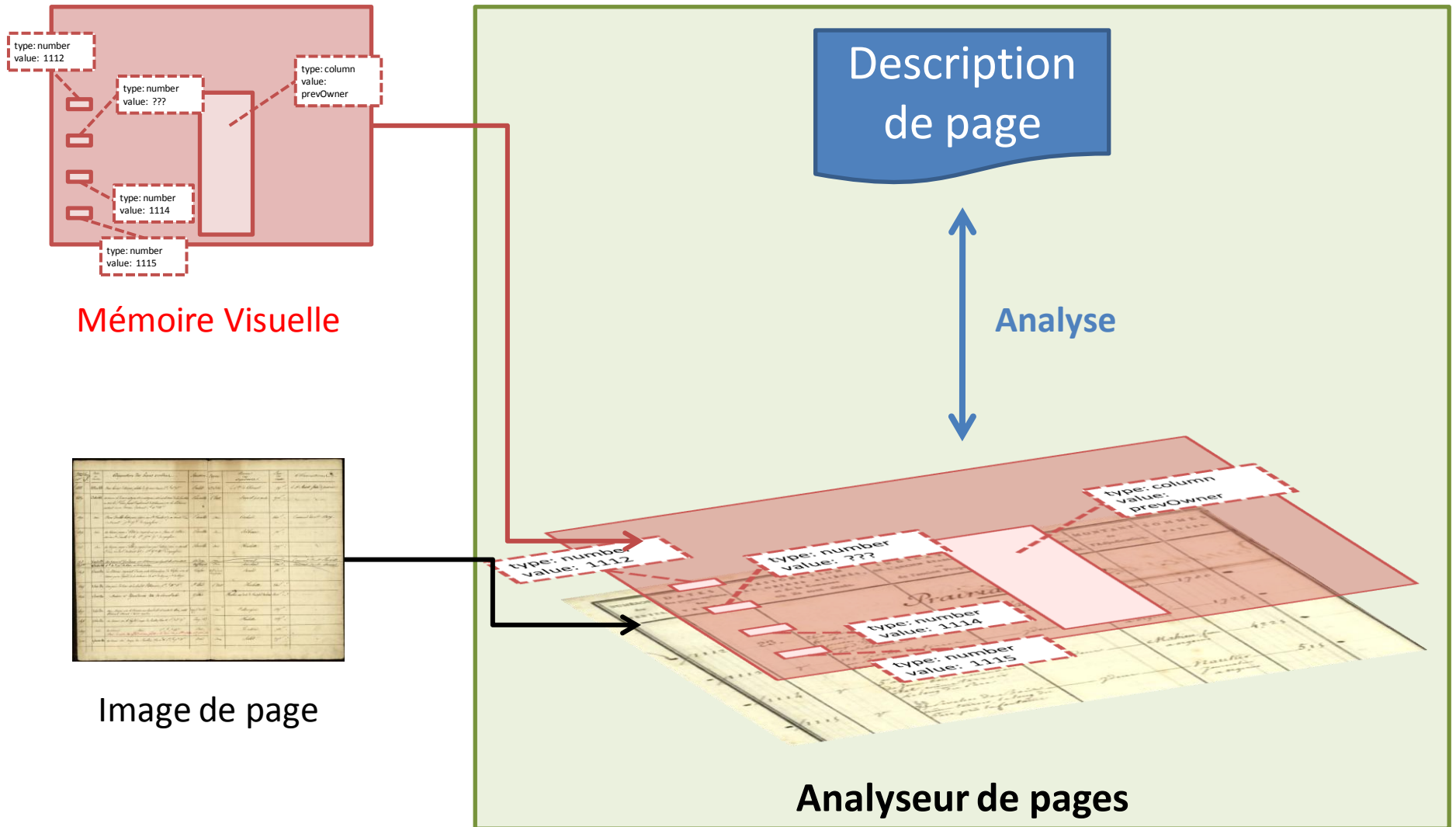
# 2/ Analyse de page itérative (1/2)

## Communiquer avec l'extérieur



# 2/ Analyse de page itérative (2/2)

Fusionner les informations (image et extérieur)



# 3/ Nouveaux opérateurs de description



## Description de page (grammaire)

- Éléments à localiser et reconnaître
- Positionnement relatif et absolu
- Autres propriétés utiles

## Comportement associé de l'analyseur

- Analyse guidée par la description
- Implémentation des opérateurs = comportement analyseur

## Nouveaux opérateurs : sémantique

- Construction illégales ou suspectes  
→ **détection d'erreur**
- Parties indépendantes de la description de page  
→ **récupération sur erreur**
- Partie de la description qui peut être impactée par une erreur / remplacée  
→ **correction d'erreur**

## Nouveaux opérateurs : implémentation

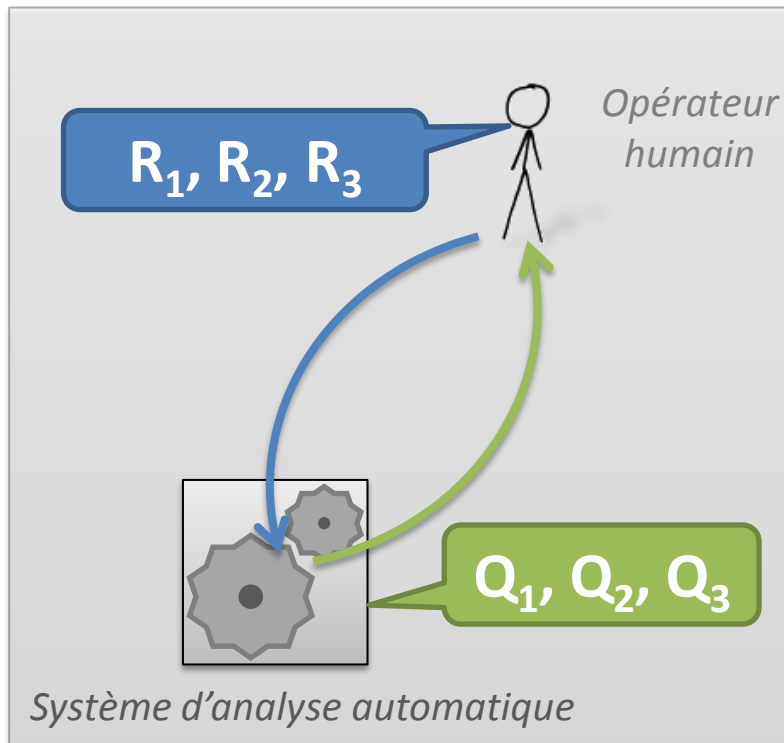
- Sauvegarde de questions
- Mise en attente de parties de l'analyse
- Remplacement de résultats intermédiaires par des données externes
- Mécanisme d'exception standard
- Gestion automatique de la communication

# Modes d'interaction proposés

## Interaction dirigée

ICDAR  
2011

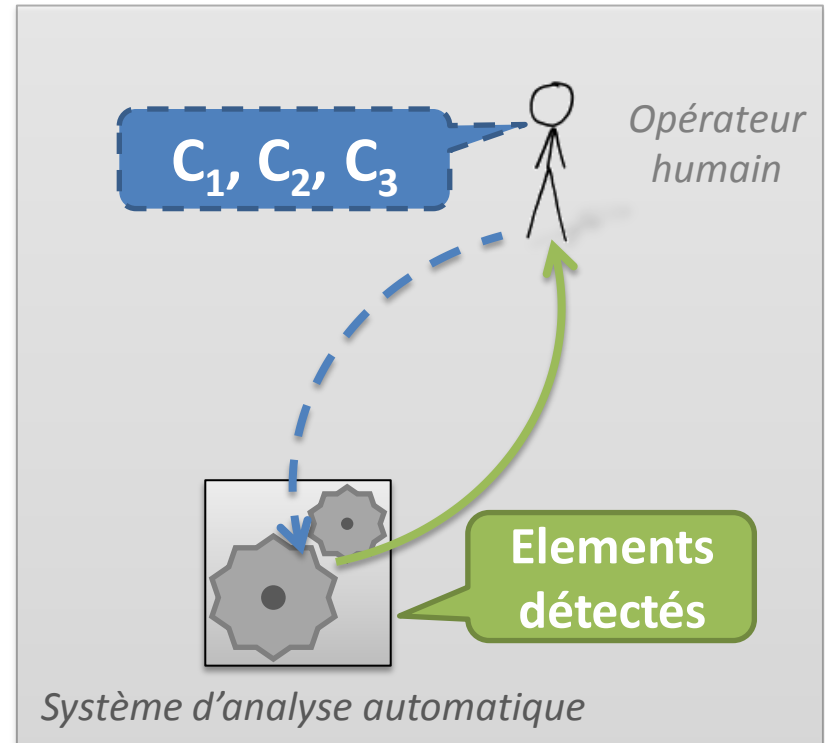
- [**sys.**] détecte erreurs **automatiquement**
- [**sys.**] pose des questions
- [**hum.**] répond à **chaque** question
- [**sys.**] utilise les réponses pour progresser



## Interaction spontanée

DRR  
2012

- [**sys.**] renvoie éléments détectés
- [**hum.**] propose des corrections
- [**sys.**] utilise informations externes (**si possible**)



Exemple 1 :

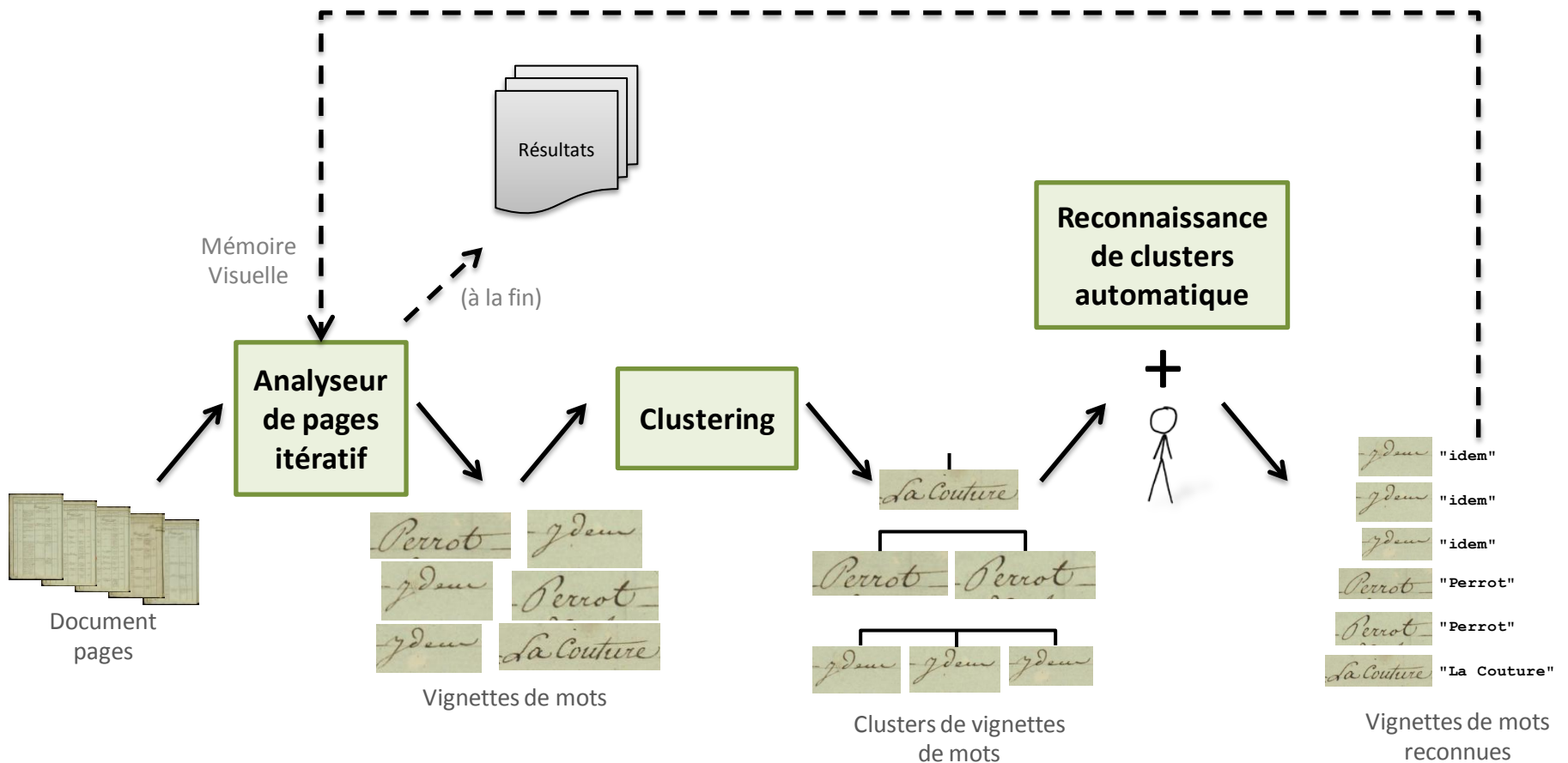
Transcription de patronymes manuscrits

**INTERACTION DIRIGÉE**

# Registres de ventes du XVIIIe siècle

NUMÉROS des VENTES.	DATES des procès-verbaux des VENTES.	DÉSIGNATION DES OBJETS ALIÉNÉS, et de la Commune où ils sont situés.	INDICATION DE L'ANCIEN ÉTABLISSEMENT, ou de l'ancien Propriétaire.	NOM de l'Adjudicataire ou de son Command.	MONTANT de l'Adjudication.	SOMME PAYÉES.	SOMME QUI RESTENT D en capital.
<u>Ferrier</u> 1791.							
7.	3	5 arpents de pré terroir deff. Germain - les arpaizon, ch. de la Boinelle	Cure deff. Germain des Arpaizon	Perrot m. de Boi à Arpaizon	5,25		
8	7.	3 arpents de terre et surcense, en 2 pièces même terroir, ch. de grand S. mâté	Religieuses deff. Cuthropes les chanteloup	Deliot culteur deff. deff. deff. à Arpaizon	2,075		
9	7.	12 arpents 84 perches 3 de terre, en une pièce même terroir, et Champêtre	7 deff.	Berson m. deff. deff. à Marolles sur le poix	10,600		
10	7.	4 arpents 24 perches de vignes, en 2 pièces même terroir, tenant au Parc deff. chanteloup, et au chemin de la Rocession	7 deff.	La Couture Bourgeois et Genard Gaimier deff. deff. deff. deff. à Arpaizon	4,550		
11	7.	2 arpents de terre, même terroir, ch. de la Ceinture	7 deff.	Loret Patronus tracteur à Arpaizon	1,725		
12	7.	3 arpents 75 perches de pré, même terroir	Picure deff. Guenand de Corbeil	Perrot m. de Boi à Arpaizon	3,300		

# Protocole



1 – Extraire et collecter les vignettes de mots de **toutes les pages**

2 – **Regrouper** les mots de pages voisines

3 – Reconnaître les clusters **avec aide humaine** si nécessaire

4 – **Réanalyser toutes les pages** pour produire des résultats structurés

# Résultats

Données : 70 pages (1206 vignettes)

**Utilisation en indexation : 20% d'erreur**

➔ Diminution relative de 28% de la quantité de travail manuel

**Utilisation en auto-apprentissage : 1% d'erreur**

➔ Gain relatif de 58% en annotations automatiques (grâce aux regroupements)



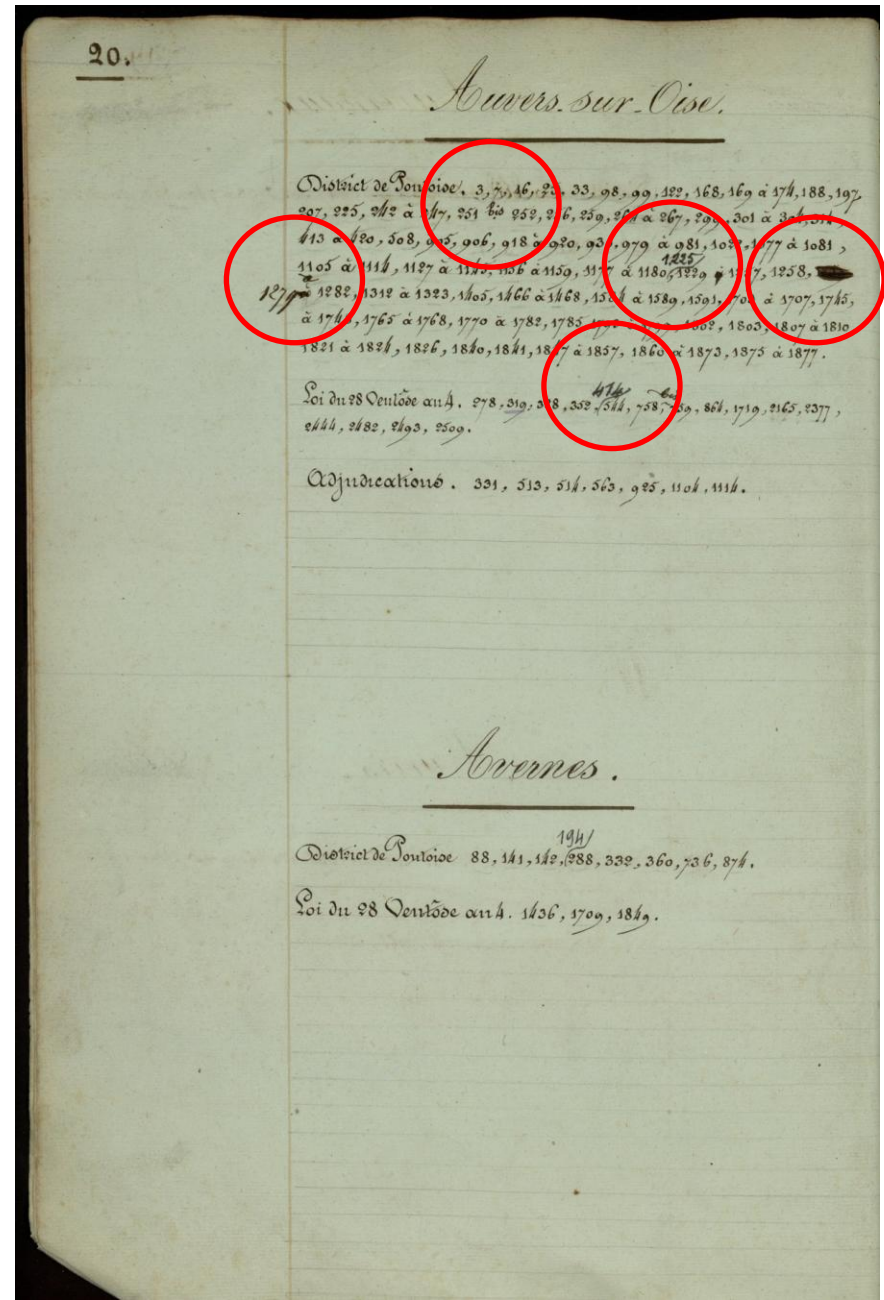
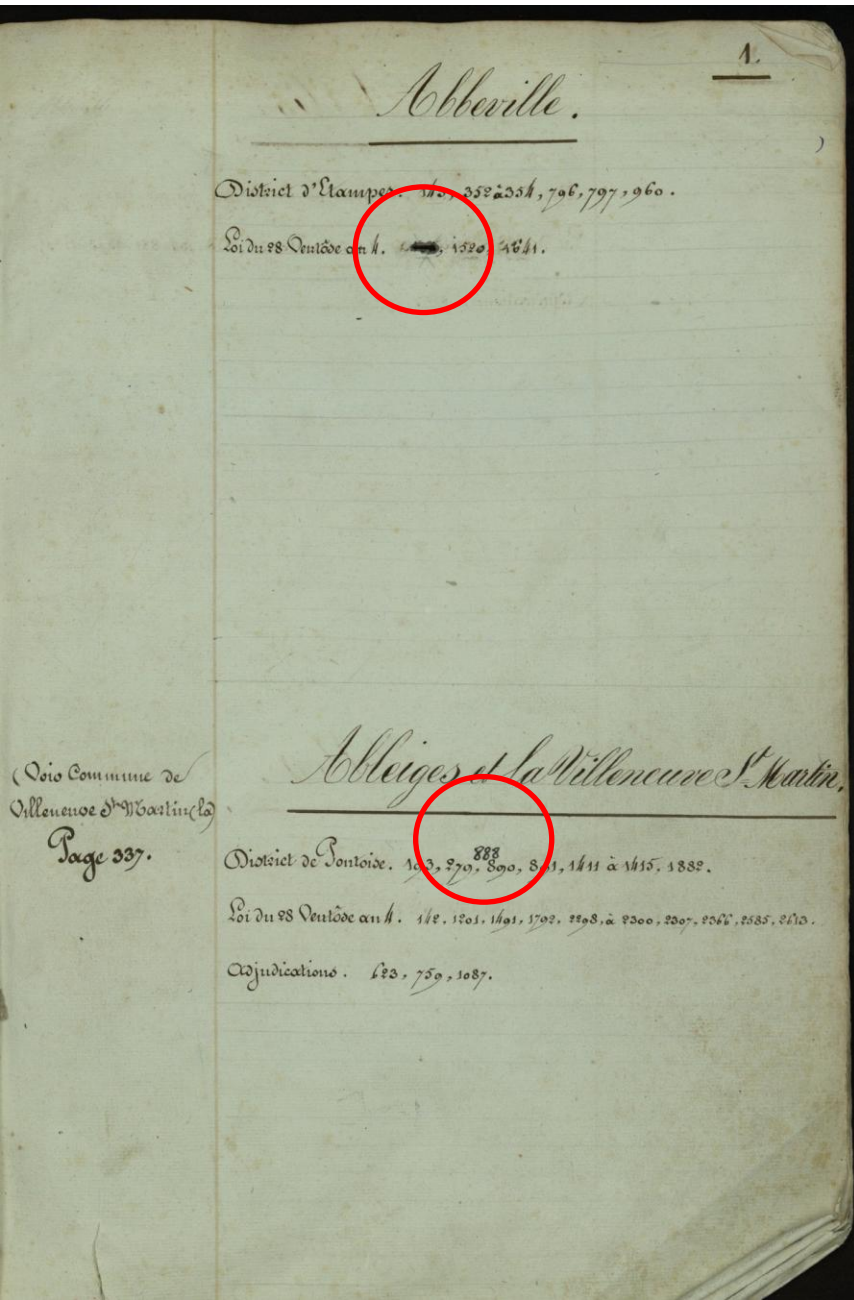
Exemple 2 :

Correction de sous-segmentation de nombres manuscrits

**INTERACTION SPONTANÉE**

# Index de ventes du XVIIIe siècle

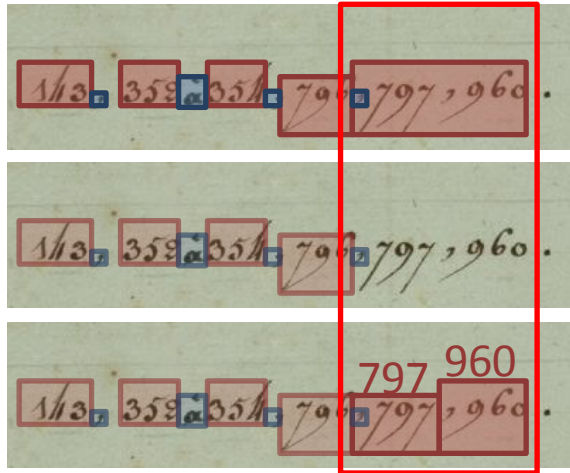
18



# Protocole

Baseline :

Correction en **post-processing**



1. Localiser les cas de sous-segmentation dans les résultats

*Comment générer la structure finale ?*

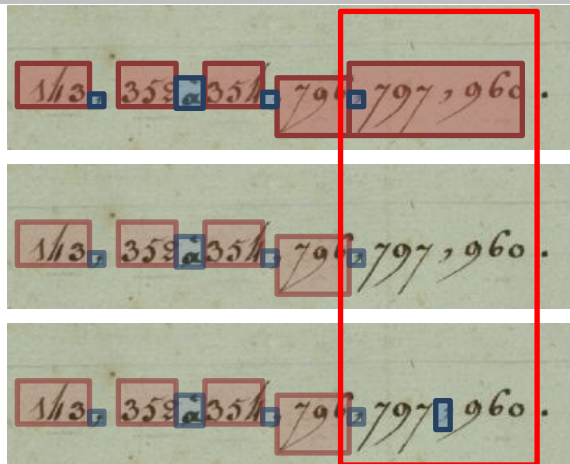
2. Nettoyer la zone

3. Ajouter une zone pour chaque nombre (et saisir valeur)

*Opération coûteuse*

Notre approche:

Correction **pendant l'analyse**



1. Localiser les cas de sous-segmentation dans les résultats

*La reconnaissance des nombres et la génération de la structure finale sont automatiques*

2. Nettoyer la zone

3. **Ajouter le séparateur manquant**  
(avec un stylet, par ex.)

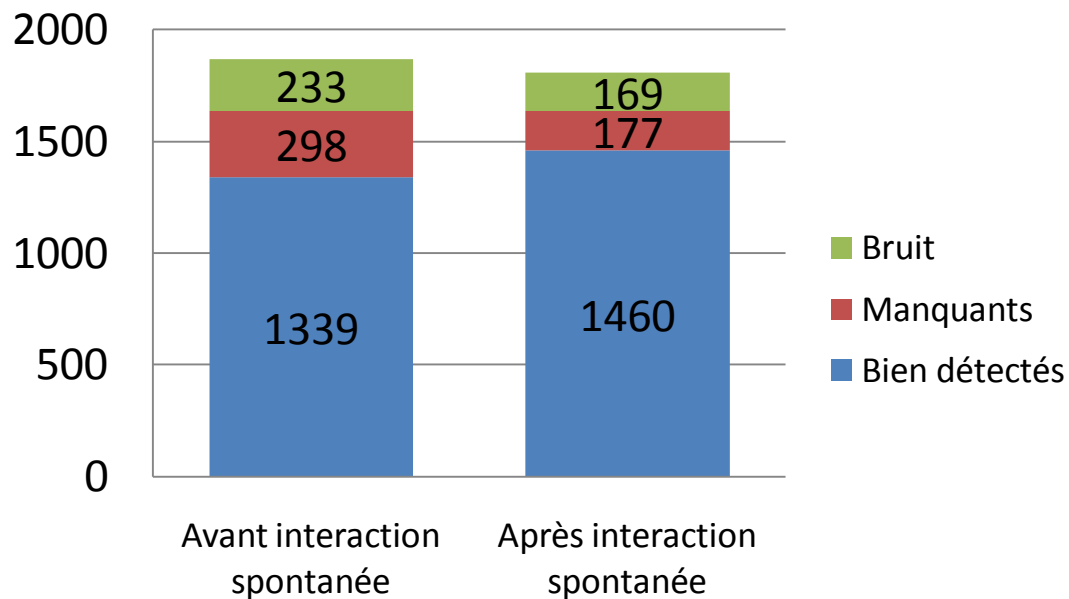
# Résultats

Données : 50 pages (1637 vignettes de nombres)

- Évaluation de la qualité de la localisation

## Variations en qualité et en coût

Variation en qualité (# zones bien loc.)



Coût pour atteindre cette qualité (# actions)

Post-processing (pas d'interaction)	121
Avec interaction spontanée	85 (-30%)

# EN RÉSUMÉ

# Interaction asynchrone efficace

- Apport d'information par l'humain doit être
  - **Asynchrone**
  - **Pendant la phase d'analyse**
- Contributions
  - **Framework** d'analyse asynchrone
  - Gestion des erreurs avec **2 modes d'interaction complémentaires**
  - **Feuille de route** pour transformer un système existant
- Résultats
  - **Diminution du coût de correction manuel** dans plusieurs situations
  - Conception **simple**, voire plus simple

# Et pour aller plus loin...

- Détails à propos des exemples présentés
  - Information échangées
  - Déroulement de l'analyse
- Evaluation du coût d'interaction
  - Exemple représentation qualité vs. coût
  - Difficulté de l'évaluation du coût
- Cas réels
  - Interactions hybrides
  - « Interaction »  $\neq$  « Interaction avec un humain »