

IA ET APPROCHES  
PARTICIPATIVES DANS LES  
HUMANITÉS NUMÉRIQUES : DE  
LA CONJONCTION À LA  
COOPÉRATION

Humanistica 2021

Olivier Aubert, Guillaume Raschia, Françoise Rubellin, Benjamin Hervy  
avec la collaboration de Julien Le Roux et Alan Le Verge

## LE CONTEXTE

Étude des théâtres sans privilèges  
au XVIIIème ( Théâtres de la Foire,  
Comédie Italienne)

- étude des pièces non-éditées (**Ciresfi**)
- études musicales (**Theaville**)
- reconstitutions en réalité virtuelle (**VESPACE**)
- étude des registres de comptes (**RECITAL**)





# DES DONNÉES POUR REPENSER L'HISTOIRE DU THÉÂTRE

## **Histoire sociale**




composition du public, placement, circulation des acteurs...

## **Histoire matérielle**

fabrique des décors, costumes, mise en scène...

## **Histoire économique**

droits d'auteurs, premiers abonnements, règles  
comptables...

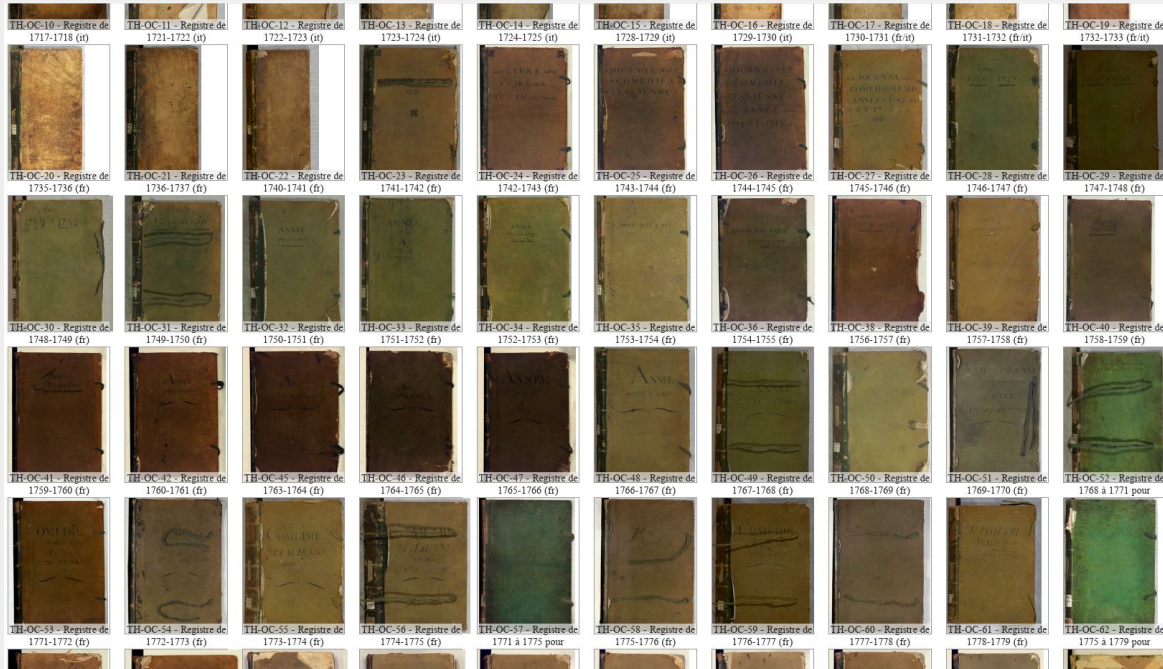


# LES REGISTRES

conservés sous les toits de la  
Bibliothèque-musée de l'Opéra



# CORPUS NUMÉRISÉ



Registres de la  
Comédie-Italienne  
de 1717 à 1794

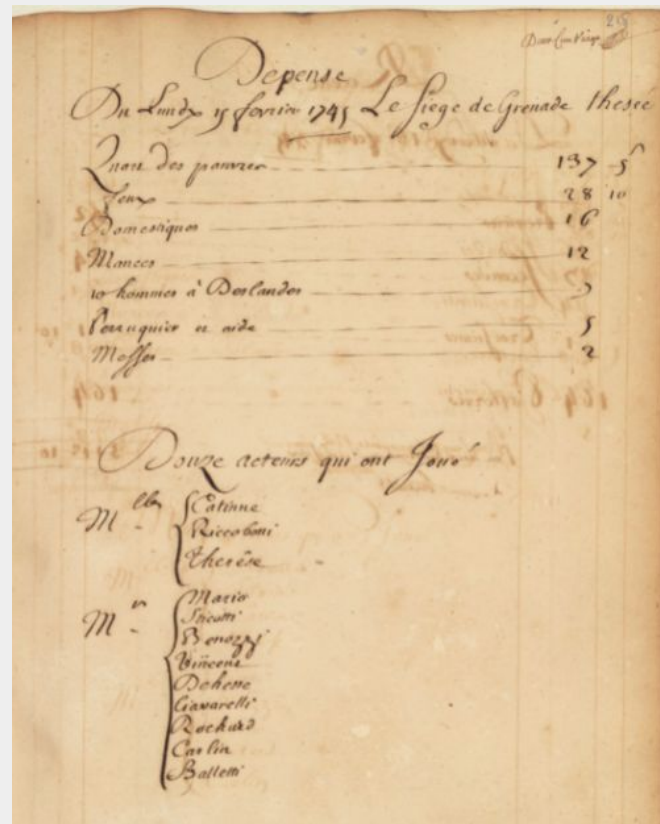
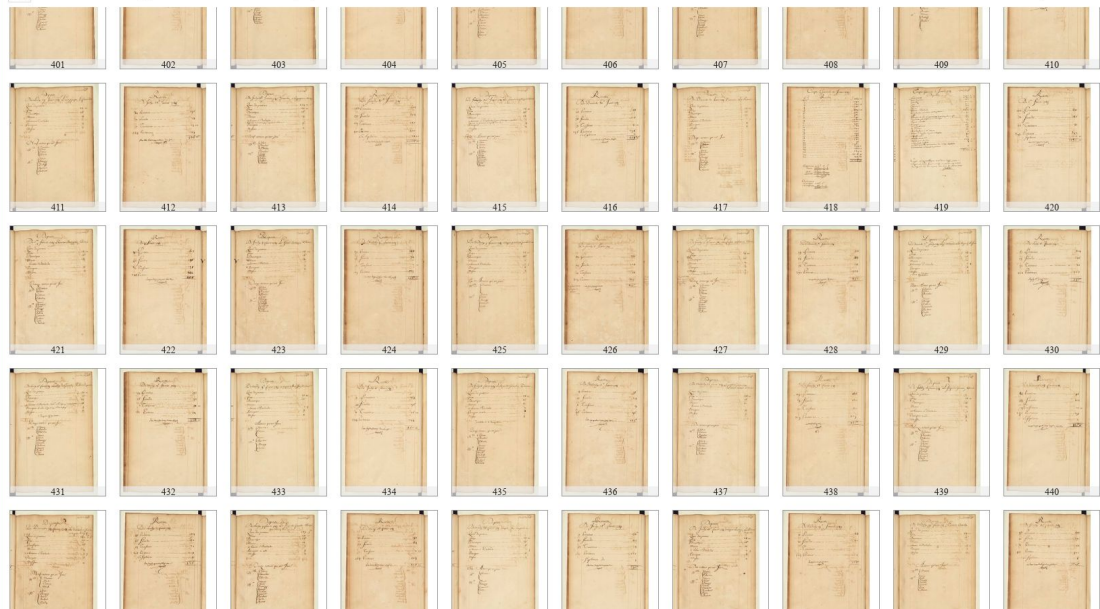
64 registres de 300 à 600  
pages

soit 26 000 pages environ

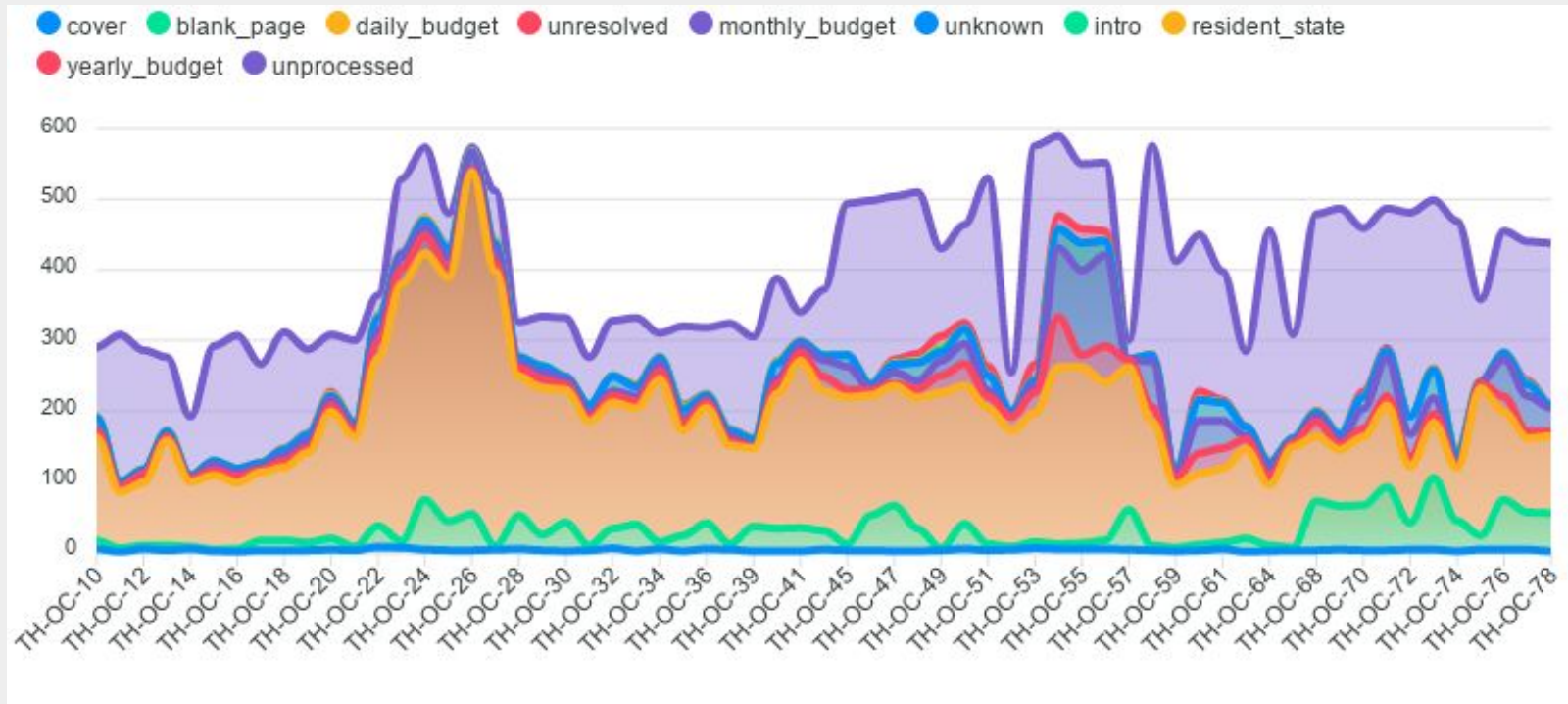
# NATURE DES INFORMATIONS

TH-OC-26 - Registre de 1744-1745 (fr)

Entrez un num.



# RÉPARTITION DES TYPES DE PAGES



# CARACTÉRISTIQUES DU CORPUS ORIGINAL

- écriture manuscrite irrégulière
- plus de 7 rédacteurs différents
- 2+ langues : italien (dialectes) et français
- système monétaire duodécimal de l'Ancien Régime :  
1 livre = 20 sous ; 1 sou = 12 deniers
- au cours du siècle :
  - évolution formelle des bilans comptables
  - évolution des règles comptables








# LE PROJET RECITAL

ANR Sep. 2014 - 2018

Du physique au numérique : deux approches initiales

Automatisation (IA) et Approche manuelle  
(crowdsourcing)







# AUTOMATISATION DU PROCESSUS PAR L'IA

Approche initiale d'utilisation des techniques d'Intelligence Artificielle

- travaux de Christian Viard-Gaudin et Harold Mouchère (équipe IVC)
- thèse d'Adeline Granet (2015-2018)

Initialement pour : découpage spatial, recherche par l'exemple, classification de page

Approche prometteuse mais se heurtant à de nombreuses difficultés techniques

- données irrégulières
  - manque de données d'exemple pour l'apprentissage
  - pas/peu de modèles adaptés
- 
- 

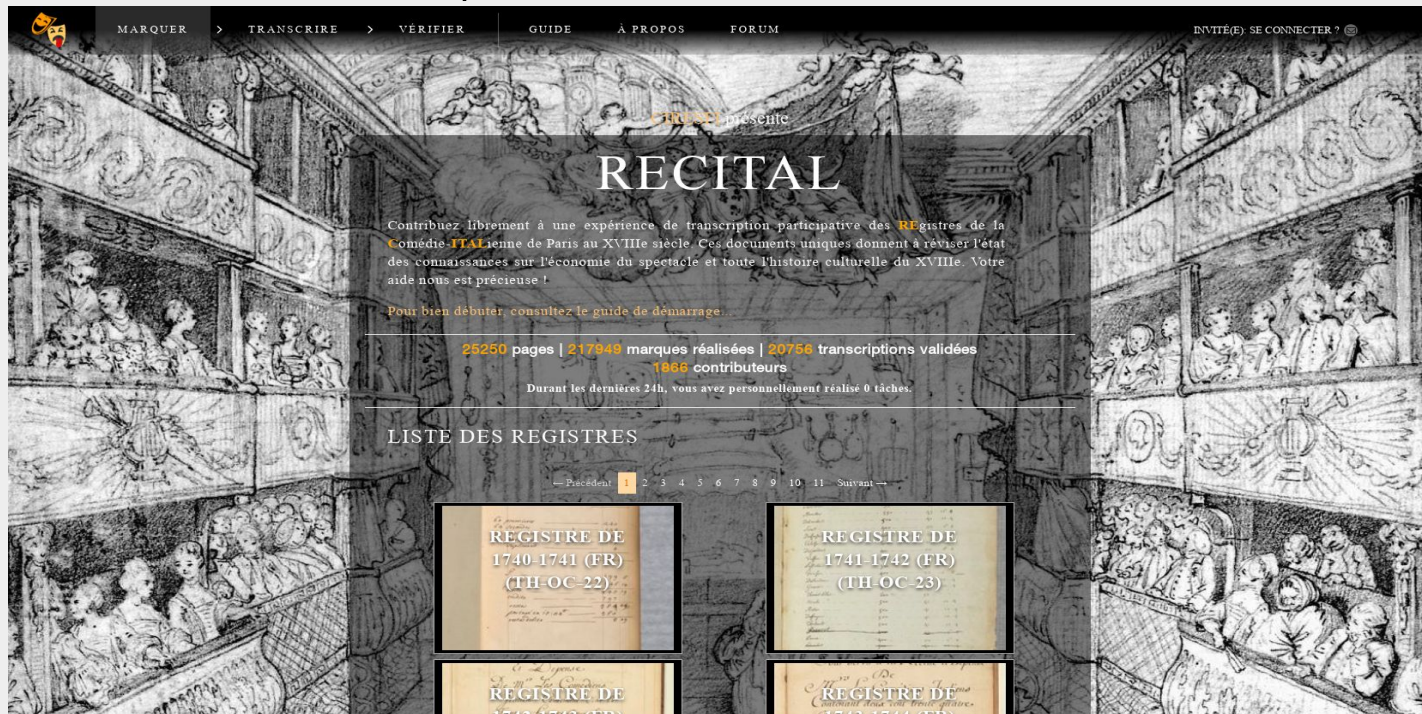
# APPROCHE CROWDSOURCING

## Approche manuelle (crowdsourcing)

- Plus laborieuse
- Mais permettant également l'appropriation du corpus et l'identification d'éléments "exceptionnels"
- Adaptée aux contenus irréguliers
- Source de données pour l'approche IA / HTR
- Principaux enjeux
  - mécanismes d'incitation (animation de communauté, ludification...)
  - enjeux de définition et d'affectation de tâche
  - enjeux de confiance dans l'information
  - enjeux de validation

# LA DÉMARCHE DE CROWDSOURCING

<https://recital.univ-nantes.fr/>



Navigation menu: MARQUER > TRANSCRIRE > VÉRIFIER | GUIDE À PROPOS FORUM | INVITÉ(E) SE CONNECTER ?

RECITAL

Contribuez librement à une expérience de transcription participative des **RE**gistres de la **Co**médie-**IT**alie de Paris au XVIII<sup>e</sup> siècle. Ces documents uniques donnent à réviser l'état des connaissances sur l'économie du spectacle et toute l'histoire culturelle du XVIII<sup>e</sup>. Votre aide nous est précieuse !

Pour bien débuter, consultez le guide de démarrage.

25250 pages | 217949 marques réalisées | 20756 transcriptions validées  
1866 contributeurs

Durant les derniers 24h, vous avez personnellement réalisé 0 tâches.

LISTE DES REGISTRES

— Précédent 1 2 3 4 5 6 7 8 9 10 11 Suivant —

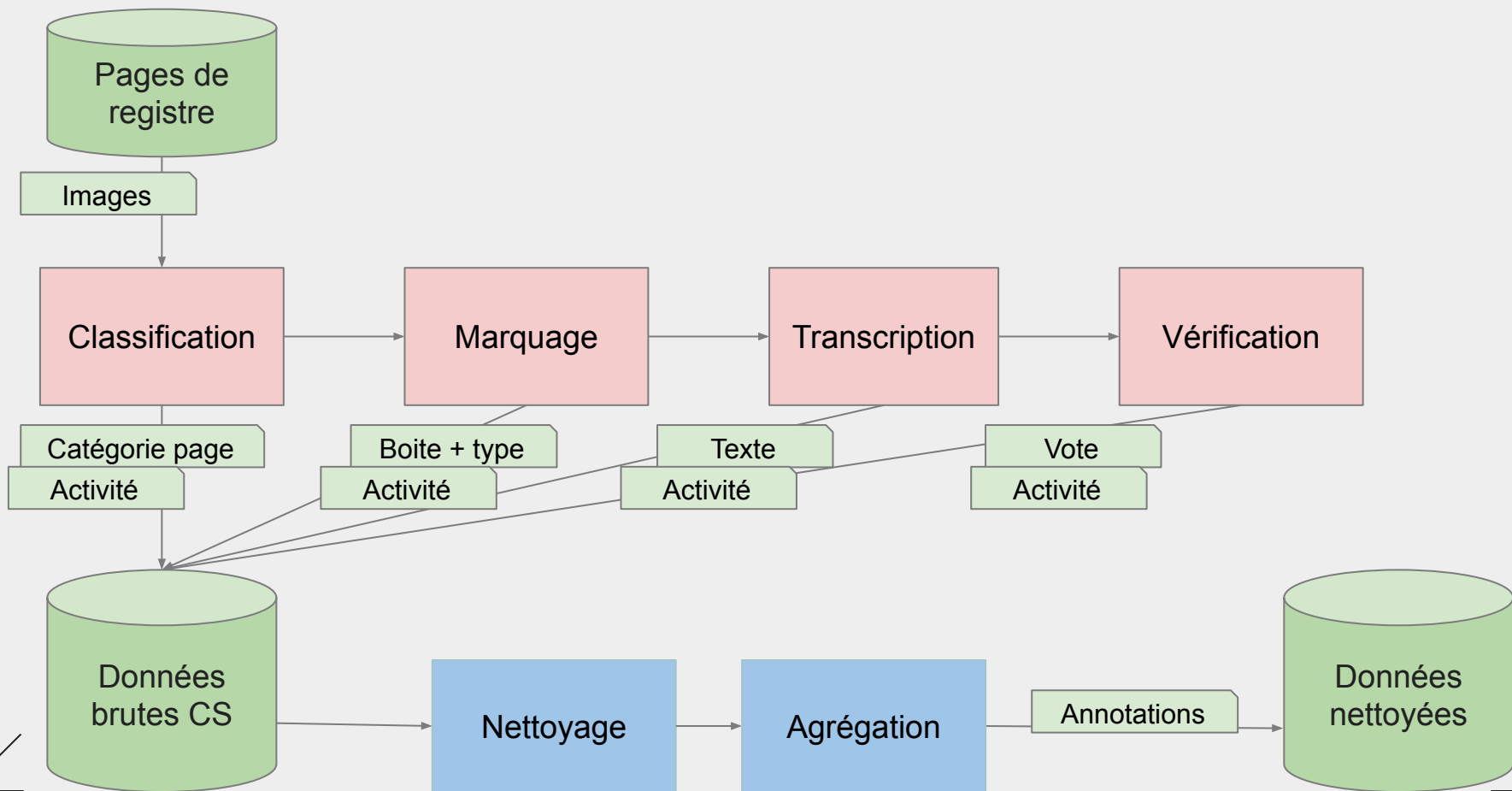
REGISTRE DE 1740-1741 (FR) (TH-OC-22)

REGISTRE DE 1741-1742 (FR) (TH-OC-23)

REGISTRE DE 1742-1743 (FR)

REGISTRE DE 1743-1744 (FR)

# WORKFLOW ACTUEL



# LES DONNÉES ACTUELLES

## Registre de 1744-1745 (fr) (TH-OC-26)

linguet

dépense

**mardi 14 avril 1744**

*arlequin sauvage le retour de tendresse*

Quart des pauvres 0

Feux 28:10

Frais de domestiques 17

Frais de mances 12

Frais de perruquier 3

2 Hommes a Deslandes

2 Hommes a Deslandes 2:10

Frais de messes 2

Nombre d'acteurs ayant joué 12

flaminia

miles

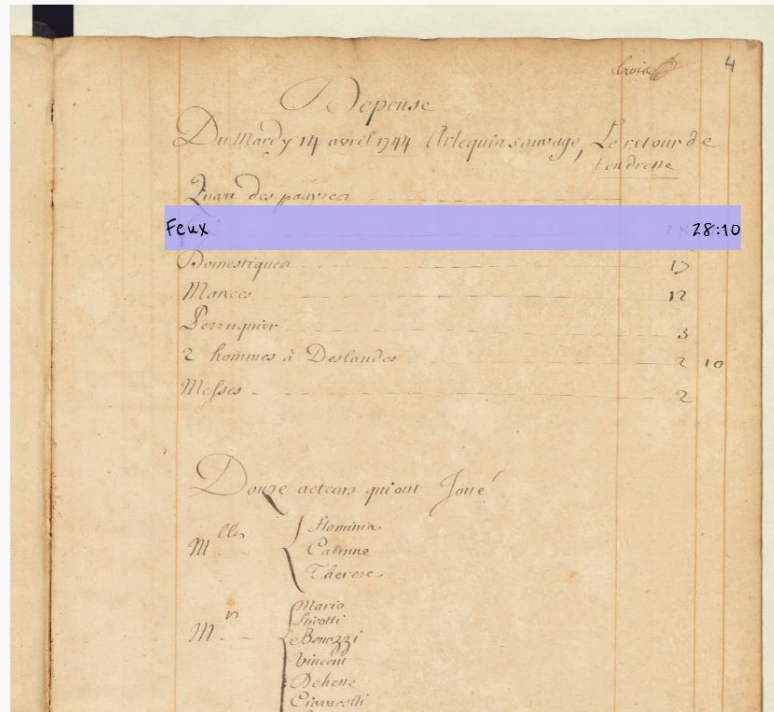
catinne

thérèse

mario

mrs

sticotti



# RÉSULTATS ACTUELS

9 catégories de pages

237 types d'annotations

834031 tâches élémentaires  
réalisées

224506 transcriptions

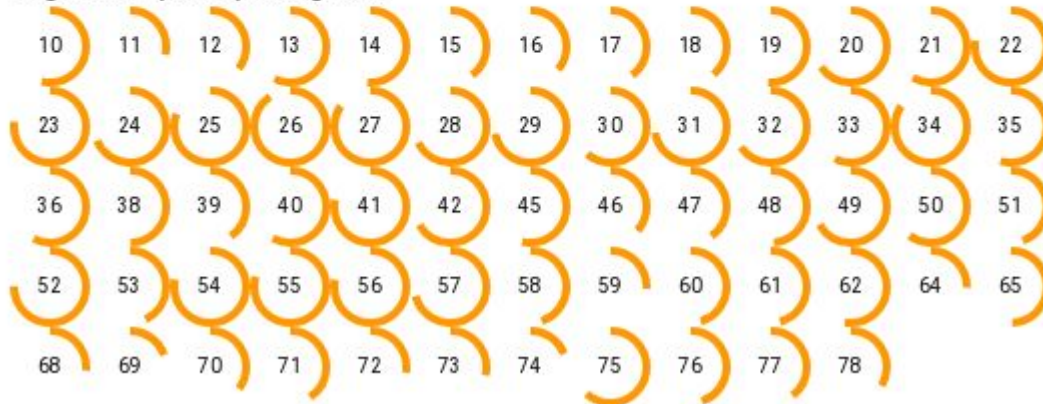
187373 annotations validées

## Avancement du marquage par registre

Nombre moyen de pages marquées sur l' ensemble des registres



### Pages marquées par registre



# VERS UNE COLLABORATION IA-HUMAIN

- Qu'entend-on par IA ?




“ find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves.” (Darthmouth Workshop proposal, 1956)

- Stage de Julien Le Roux et Alan Le Verge (étudiants Polytech)  
Étude de différents scénarios de collaboration possible entre humain et machine dans le crowdsourcing





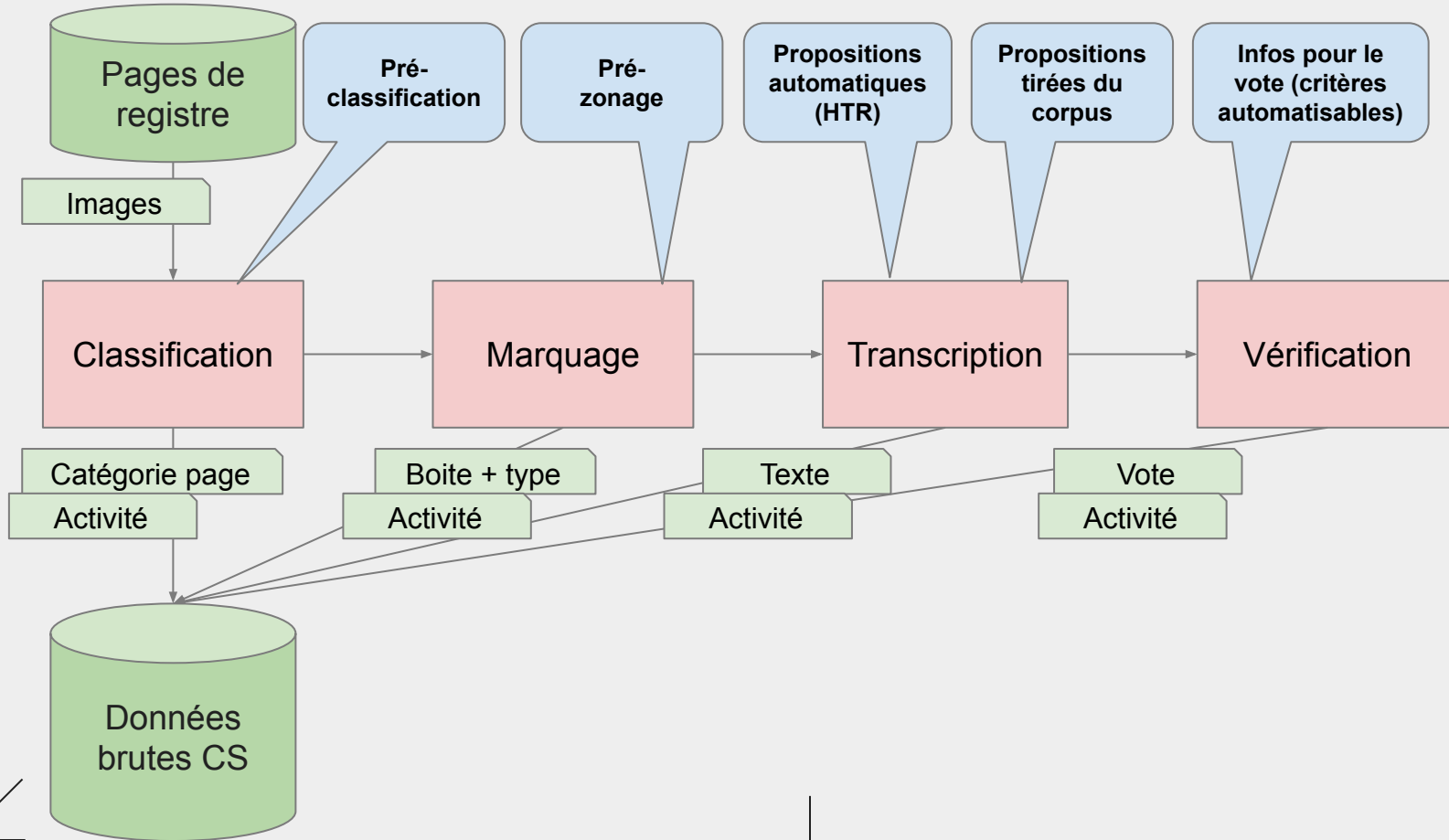
## 3 AXES DE COLLABORATION

- L'IA comme assistant au participant
  - L'IA comme assistant à l'expert/concepteur du système
  - L'IA comme utilisateur virtuel
- 
- 
- 

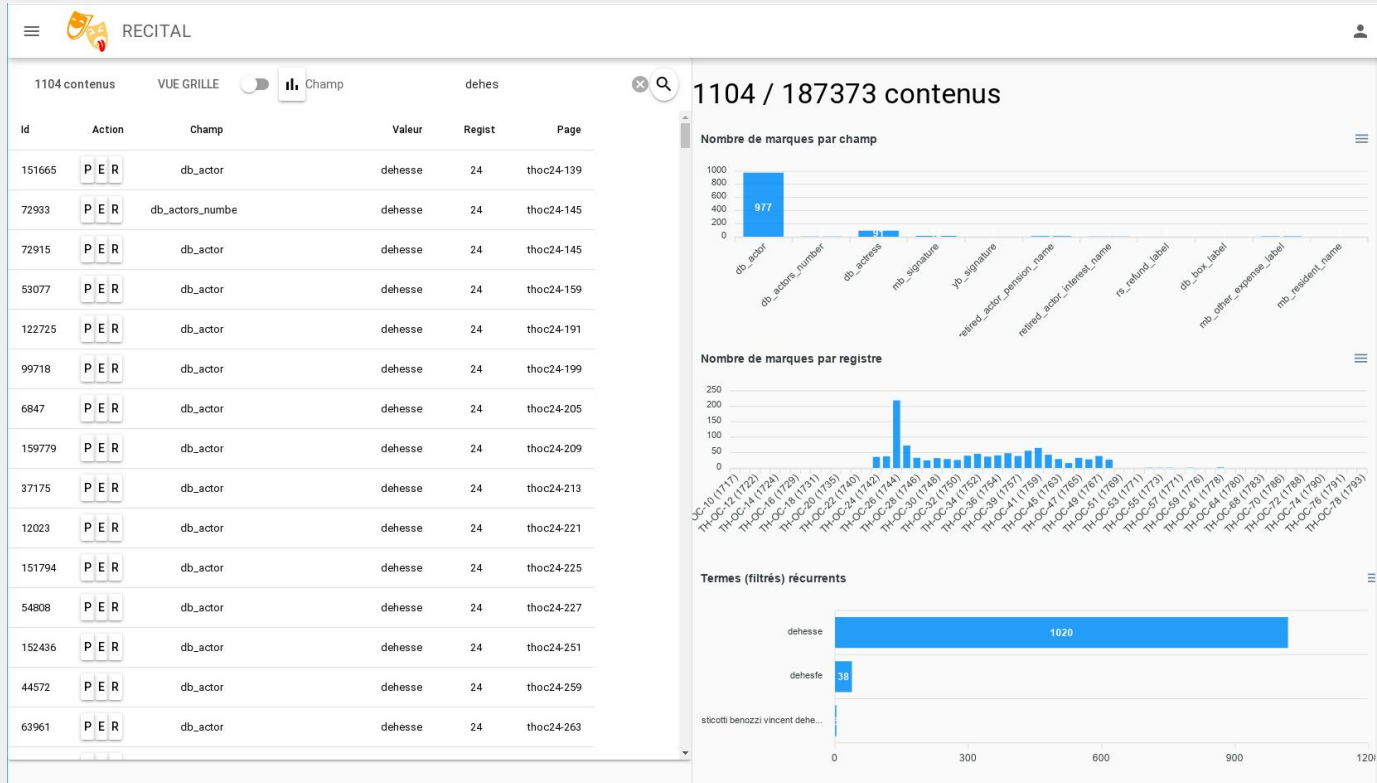
# L'IA COMME ASSISTANT AU PARTICIPANT

- Intégration de l'IA dans l'interface et les outils proposés aux participants
- Fournit des données pré-saisies, que l'utilisateur peut corriger et valider
- Approprié pour des tâches basiques (zonage par exemple)
- Si l'IA est trop efficace : risque d'ennui (uniquement validation)
- Si l'IA est trop peu efficace : contre-productif (évaluation et correction de données invalides plus long que saisie directe de données)
- Intégration à l'interface
  - soit des outils génériques externes
  - soit des développements spécifiques dans l'interface

# ASSISTANCE AU PARTICIPANT



# EXEMPLE : DASHBOARD





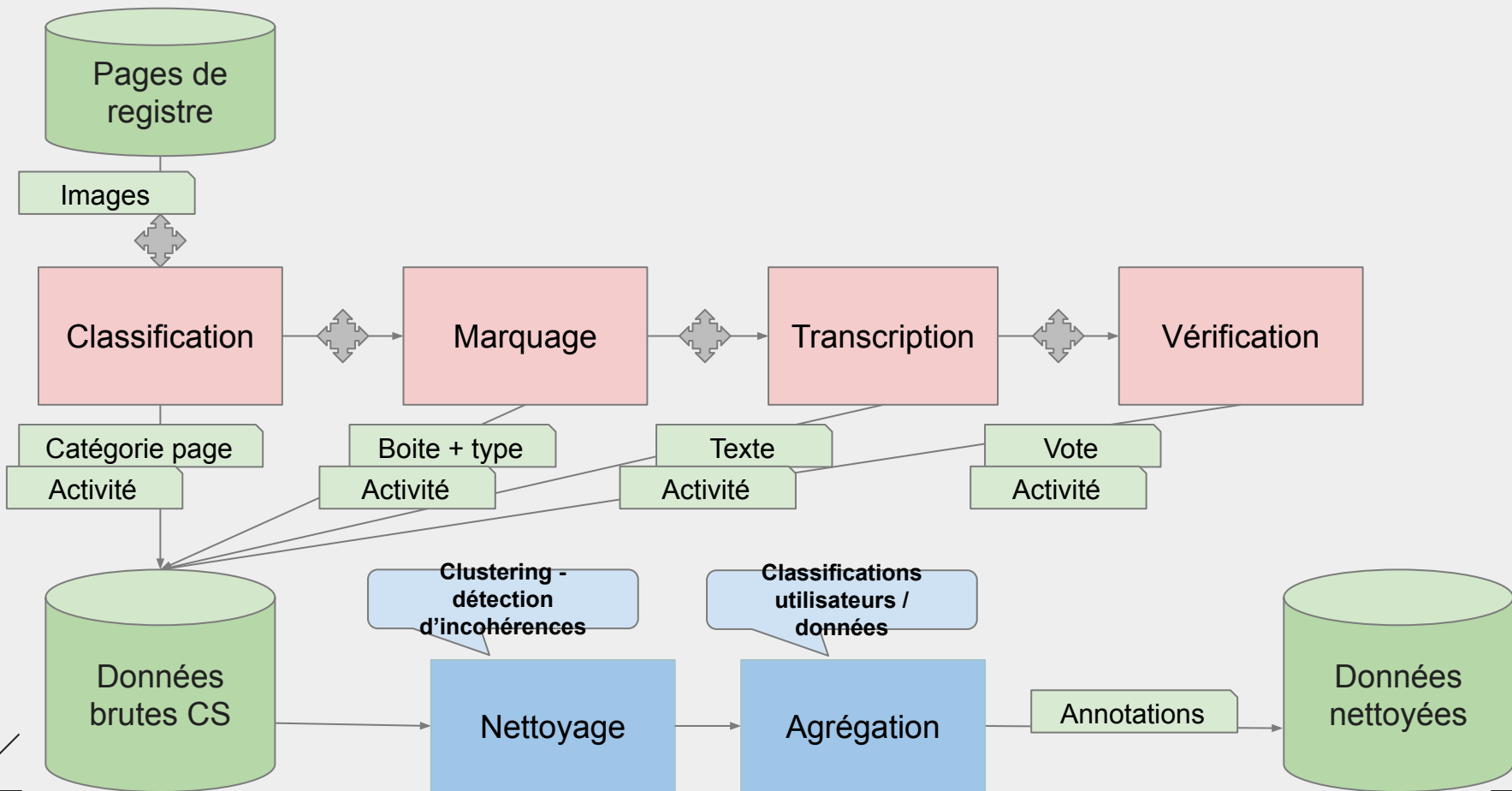
# L'IA COMME ASSISTANT À L'EXPERT/CONCEPTEUR DU SYSTÈME

- Nettoyage des données (automatisation: regexp, NLP, IA...)
- Évaluation / détection d'incohérences (classification, résolution d'entités...)
- Évaluation de l'activité utilisateur et de ses contributions
  - Classifications/profils
  - En amont pour l'orientation de l'affectation des tâches
  - En aval pour la pondération des résultats



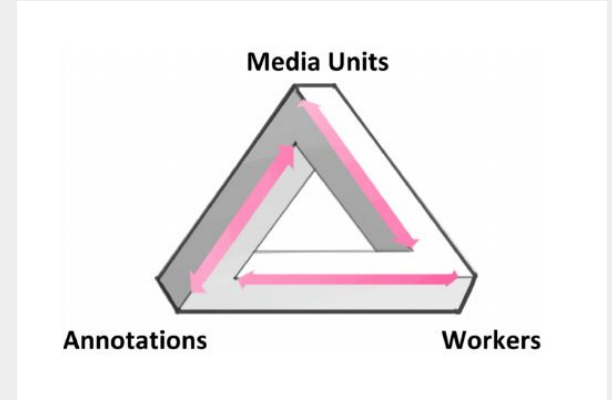
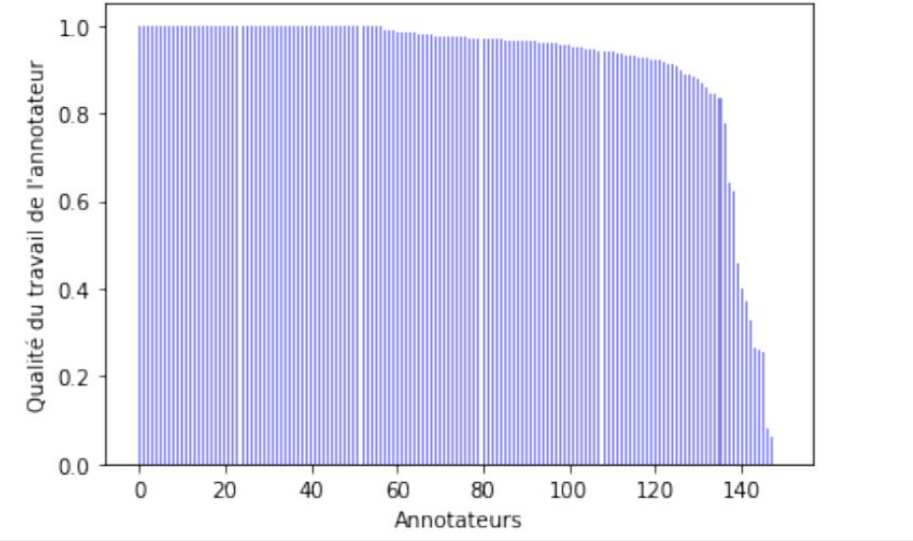
# ASSISTANCE À L'EXPERT

Affectation de tâches



# EXEMPLE - PLATE-FORME CROWDTRUTH

Histogramme représentant la qualité du travail de chaque annotateur



Projet étudiant Yann SALMON, Dorian MASSAMBA et Samuel MOUTINHO

# L'IA COMME UTILISATEUR VIRTUEL

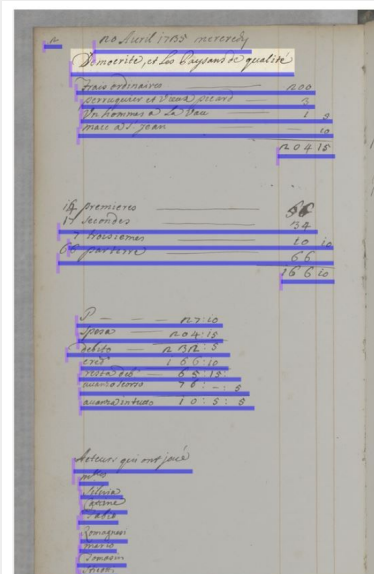
- Introduction d'un agent autonome qui va "simuler" un utilisateur dans son interaction (dans la mesure des possibilités de l'outil/interface)
- Production de données avec le même statut qu'un utilisateur
- Statut "conjoint" plutôt que superviseur



# TESTS TRANSKRIBUS / ESCRIPTORIUM

Expérimentation : on fait  
“apprendre” à un utilisateur  
virtuel sur la base des  
marques déjà transcrites puis  
on lui demande de participer  
à la transcription

- conversion des données  
crowdsourcées en ALTO
- apprentissage de  
modèle



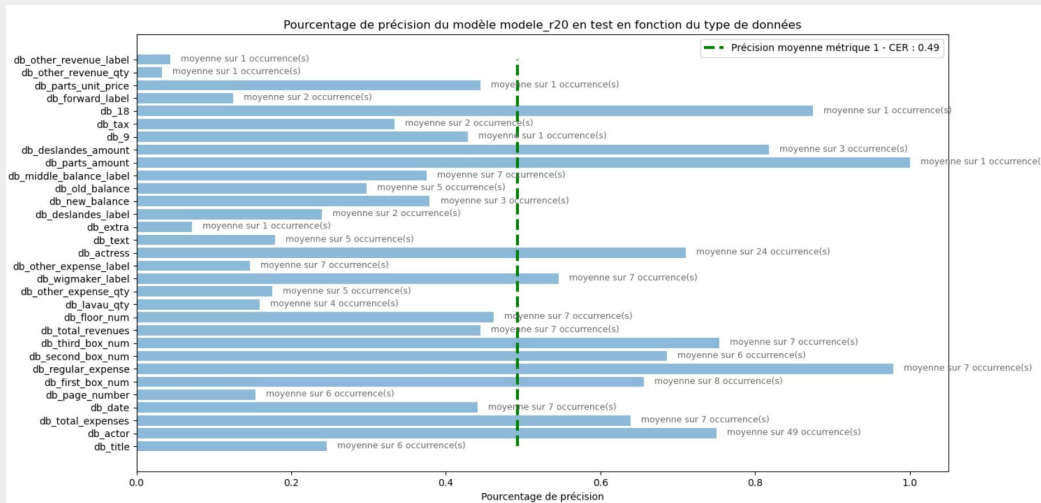
2 mercredi 20 avril 1735  
democrite et les paysans de qualite  
200  
perruquier et vieux picard 3  
1:5 1 un homme à la vau  
10 mance à saint jean  
204:15  
17 34  
7 10:10  
66 66  
166:10  
27:10  
spesa 204:15  
232:5  
credito 166:10  
resta debito 65:15  
avanza scorso 76:5  
10:5:5  
acteurs qui ont joué  
mesd  
silvia  
cattine  
fabio  
romagn  
maric  
tomasin  
sticott

1 mercredi 20 avril 1735  
2 2  
3 democrite et les paysans de qualite  
4 17 34  
5 7 10:10  
6 66 66  
7 200  
8 perruquier et vieux picard 3  
9 1:5 1 un homme à la vau  
10 10 mance à saint jean  
11 204:15  
12 166:10  
13 27:10  
14 232:5  
15 10:5:5  
16 silvia  
17 mesdemoiselles  
18 cattine  
19 fabio  
20 romagnesi  
21 mario  
22 tomasin  
23 sticotti  
24 benozzi  
25 vincent

# TESTS TRANSKRIBUS / ESCRIPTORIUM

Comment faire correspondre au mieux les idées avec les capacités des outils ?

- Typologie des boîtes non prise en compte nativement
- Mais est corrélé à la qualité de la reconnaissance
- Adaptation de modèle de segmentation (modèle "unitaire" vs fragment)





# CONCLUSION

Champ défriché, en cours d'exploration

Besoins identifiés

- ingénierie
- évaluation
- interfaçage (API, GUI)

Réflexions généralisables à d'autres contextes de crowdsourcing?

