

Correction des données : retour d'expérience sur la plate-forme RECITAL de transcription participative

Benjamin Hervy, Pierre Pétillon, Hugo Pigeon, Guillaume Raschia

Post-doc, ANR CIRESFI
LS2N - UMR CNRS 6004
Équipe DUKE



ANR CIRESEFI (2015-2018)

- {C}ontrainte et {I}ntégration : pour une {RE}évaluation des {S}pectacles {F}orains et {I}taliens
- Analyse de documents autour de la Comédie-Italienne et des théâtres forains
- Plusieurs corpus à l'étude dont les **registres comptables de la Comédie-Italienne (1717-1794)**

Deux sous-projets (sciences des données) menés en parallèle :

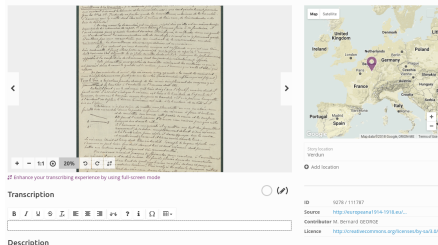
- Reconnaissance automatique d'écriture manuscrite (A. Granet, doctorante LS2N)
- **Projet RECITAL de transcription collaborative** (crowdsourcing) pour pallier les limites

Crowdsourcing

Collecte de données historiques : crowdsourcing



FRBDIC- 11 Histoire de Louis Carette, capitaine d'infanterie puis instructeur détaché aux Armées américaines, item 4

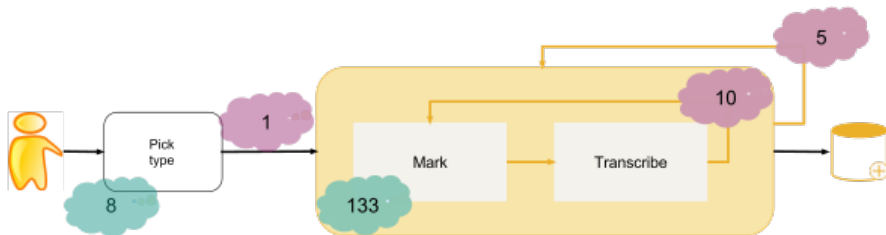


- 1 Projets de crowdsourcing orientés édition (*Transcribathon*, *Testaments de poilus*, *Transcribe Bentham*)
- 2 Projet RECITAL (catégorisation des informations en plus de la transcription + évaluation collaborative)

RECITAL

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

Processus de crowdsourcing



Quelques caractéristiques :

- 25250 documents (63 registres), 8 types de documents, 133 catégories tous types confondus
- 2 propositions identiques nécessaires pour la validation
- Vote à la majorité (75%) si désaccord

Verrous techniques et scientifiques

- Qualification et nettoyage des données (*Rham et Do, 2000*)
 - ▶ Recherche de consensus par la foule
 - ▶ Résolution des conflits (complexité + interprétation des instructions)
 - ▶ Résolution d'entités (variations dans les données)
 - ▶ Alignement : cas des titres de pièces
- Transformation de modèle (*Fagin et al., 2005*) :
 $\{\text{Action, Utilisateur}\} \Rightarrow \{\text{Marque, Transcription}\} \Rightarrow \{\text{Entité métier}\}$
- Filtrage, exploration et résumé statistique

Du bruit dans les données

8 Août Lundi		17
6 ^e Trappe (Confiance, et Layettes imprimees)		
Trois ordinaires		2 00
Portraquer		3
Motivo		2
15 ^e Domestiques		7 15
Propre		1 5
Mante au balconnier d'argent pour crayonné		1
		2 15
8 premiers		
6 seconds		1 32
3 troisiemes		5 2
20 porteres		15
		2 00
		3 19
4 ^e 2 ^e passeres		
1 ^{er} pass		7 9
2 ^e pass		2 13
18 ^e pass		2 94 15
		3 5 9
Debit		
credit		2 98 - 9
debit		3 9
debit		2 1
debit		2 99 18 6
debit in banco		4 970 12 9

[illegible]

Du bruit dans les données

Crowdsourcing et choix de conception

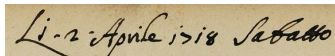
- Profil des participants
- Recherche de consensus (vote à la majorité)
- Mode de transcription libre
- Stratégie d'affectation des tâches

Corpus original

- Changement de scripteur
- Nature des comptes et granularité de l'information
- Évolution des règles comptables
- Erreurs dans le texte original

Résolution de conflits

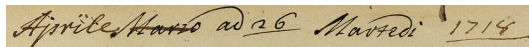
Cas faciles :



Li-2 Aprile 1718 Sabato

Proposition	Votes
02/04/1718	1
samedi 02/04/1718	4

Cas plus difficiles :



Aprile Martedì ad 26 Martedì 1718

Proposition	Votes
mardi 26 avril 1714	2
26/04/1718	2

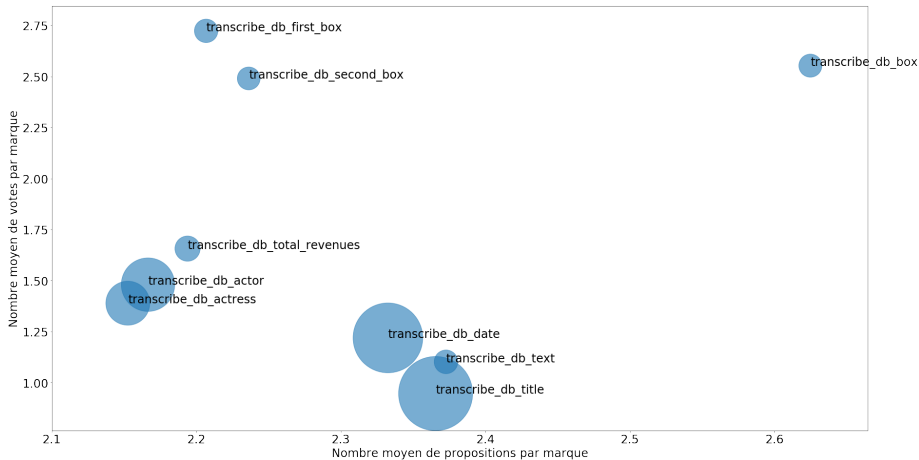
- Qu'est-ce qui rend un vote difficile ?
- Y a-t-il des types de données qui posent plus problème que les autres ?
- Pourrait-on facilement éviter certains conflits et limiter les efforts à fournir pour les travailleurs ?

Résolution de conflits

Méthodologie :

- ➊ Statistiques générales sur les données existantes
- ➋ Identification des types de données les plus problématiques : titres, dates, valeurs monétaires
- ➌ Analyse des causes de divergence pour chaque type
- ➍ Définition d'heuristiques basées sur les connaissances expertes
- ➎ Correction des données et résolution du conflit le cas échéant

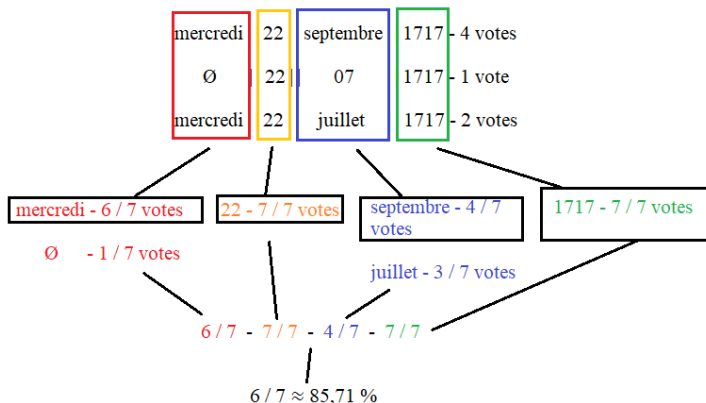
Résolution de conflits



(577 marques *transcribe_db_title*)

Résolution de conflits

Exemple sur les dates



Algorithme similaire pour les valeurs monétaires (segmentation *livres : sous : deniers*)

Résolution de conflits

Bilan de l'approche de résolution des conflits (échantillon de jan. 2018)

<i>Type</i>	<i>Marques doublées</i>	<i>Marques vote</i>	<i>% conflits résolus</i>
Dates	815	511	80
Titres	984	577	80
Valeurs monétaires	1509	456	50

- Marques doublées : marques ayant fait l'objet de 2 transcriptions
- Marques vote : marques dont la transcription fait l'objet d'un vote

Résolution d'entités

Raton Et Rosette, Les amours de Bastien, et
Les amusemens champêtres, Ballet. &c.

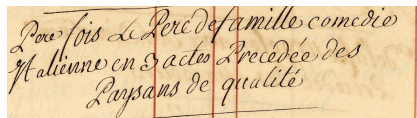
Les Corsaires
Les pêcheurs,
Les amusemens champêtres. } divertissemens
et
La Cavalcade

La Vie est un Songe,
L'impromptu des acteurs. et
Les amusemens champêtres divertissemens

Résolution d'entités

L'identification (segmentation) des titres de pièces nécessite une connaissance du domaine : alignement *a posteriori* à partir d'une chaîne globale

Alignement des titres de pièces



⇒ Sources primaires (*Gueulette*) et secondaires (*Repertorio*, etc.)

- Le père de famille
 - Paysans de qualité
-
- Modèle N-Gram et position dans la chaîne
 - Calcul de similarité avec liste de références et % de couverture de la chaîne originale
 - Amélioration : clusters de titres avec variantes plutôt que liste

Résolution d'entités : exemple

- ❶ *Première fois Le Père de famille comédie italienne en 3 actes Précédée des Paysans de qualité*
- ❷ Découpage en CharNGrams ($n = 3$) et pre-processing
- ❸ Passe n°1 $\Rightarrow$ [('Paysans de qualité', 0.44), ...]
- ❹ Passe n°2 : Recherche de la position des titres retenus en fonction de la similarité maximale et itération sur la chaîne réduite
- ❺ Stop quand la chaîne initiale est couverte à plus de 75%

$\Rightarrow$ [('Paysans de qualité', 1.0), ('Père de famille', 1.0), ('Ancienne Comédie Italienne, ou les Ombres parlantes', 0.54)]

Résultats

- Échantillon d'évaluation ($N = 150$)
- Rappel : 91%
- Précision : 92%

Exploration des données

Outils d'exploration et d'analyse : transparence des traitements opérés

Soirée du 1752-08-31

Pièces jouées

Titre	Confiance	Provenance
Heureux stratagème	1	record_linkage
Quatre âges en récréation	0.867527617235709	record_linkage

Acteurs présents

Nom
Catherine
Thérèse
Favart

Informations budgétaires

Total des recettes : 764

Type	Montant	Quantité
transcribe_db_first_box	76	304
transcribe_db_second_box	73	146
transcribe_db_third_box	10	15
transcribe_db_four	298	298
transcribe_db_extra	1	null
transcribe_db_total_revenues	764	null

Total des dépenses : undefined

Type	Montant	Quantité
transcribe_db_five	764	null

L'Heureux Stratagème et Les Ages en récréation			
		mefes	2.
76. Premiers	304.	Jeux	2916
73. Secondes	146	Domestiques	17.
		Mances	
10. Troisiemes	15	Portuquier	3.
		Paye a la Sardo	33.10
298. Partieres	298		
1. Supplement	1.		
Non payé aux fournisseurs		764.	
Acteurs qui ont joué			
		M ^{lles} Catherine	
		Thérèse	
		Favart	
		Maria	

- Disposer d'une BDD fiable, documentée et fidèle au corpus original augmentée de mécanismes d'exploration et d'analyse
- Pallier les limites des techniques de reconnaissance automatique d'écriture par le crowdsourcing
- La nature du corpus et les choix de conception de la plateforme induisent des problématiques de qualité de données
- Les techniques de correction des données permettent de résoudre 80% des conflits pour les dates et les titres, 50% pour les valeurs monétaires (ajout d'heuristiques supplémentaires)
- Les données issues de ces traitements serviront de jeu d'entraînement pour les modèles d'apprentissage (et inversement)

Au 25/05/18, RECITAL compte env. 600 contributeurs pour 100k tâches réalisées

Merci de votre attention (et bon appétit) !

benjamin.hervy@univ-nantes.fr

guillaume.raschia@univ-nantes.fr