

# RECITAL

## Combining crowdsourcing and expertise in Digital Humanities

FOSDEM - February 2021

Olivier Aubert

[www.olivieraubert.net](http://www.olivieraubert.net) - [contact@olivieraubert.net](mailto:contact@olivieraubert.net)

in collaboration with Françoise Rubellin (LAMO - Univ. Nantes)  
and Guillaume Raschia (LS2N - Univ. Nantes)

# Context: Comédie-Italienne history

RECITAL: Registres de la Comédie-Italienne

## des Théâtres de la Foire et de la Comédie-Italienne

Literature and history lab aiming at studying fairground theaters and Italian Comedy in Paris around XVIIIth century

### Multiple approaches

- edition and study of unedited plays (**Ciresfi**)
- musical studies and database (**Theaville**)
- VR reconstitution of no longer visible theatres (**VESPACE**)
- accounting registers study (**RECITAL**)

# What are we expecting to learn from accounting registers?

- performed plays titles
- sold tickets (by category)
- expenses (taxes, accessories, musicians)
- theatre accounting

We want to learn generalities, but also find out exceptional things/events



# Data to re-think theatre history

## **Economic history**

author's royalties, subscriptions,  
accounting rules...

## **Social history**

audience composition, placement,  
involved actors...

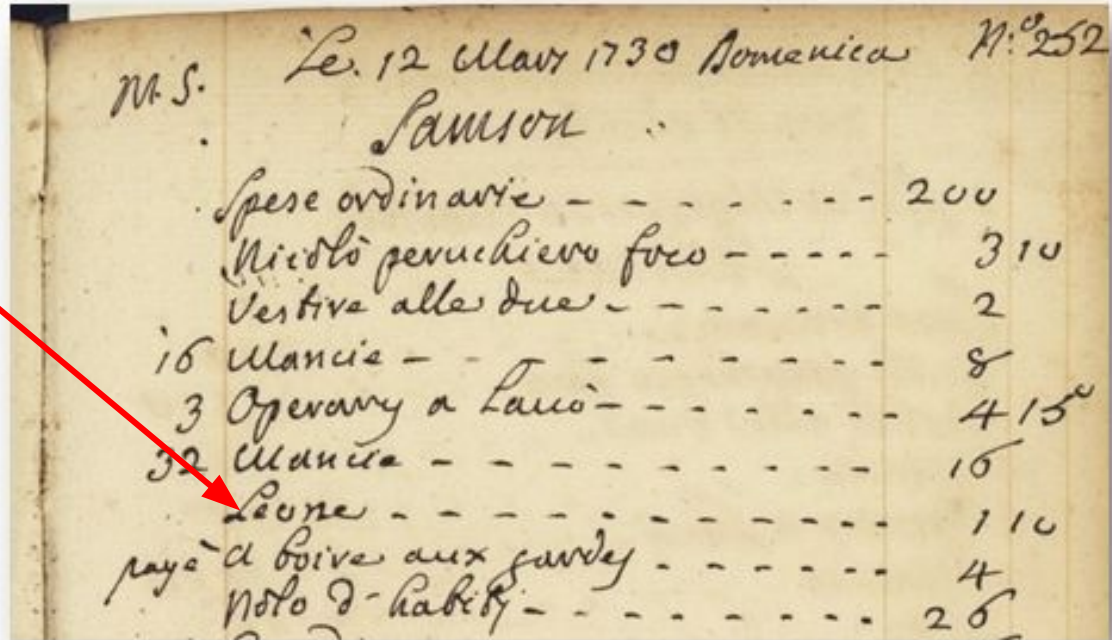
## **Hardware history**

sets making, costumes...



# Historical investigations...

Leone : Lion  
(a turkey actually)



M.S. Le. 12 May 1730 Romanica N.º 252  
Samson

Spese ordinarie - - - - -	200
Niccolò peruchiero foro - - - - -	310
Vestiva alle due - - - - -	2
16 Mancie - - - - -	8
3 Operarii a Lario - - - - -	415 <sup>0</sup>
32 Mancie - - - - -	16
Leone - - - - -	110
paye à boire aux gardes - - - - -	4
nolo d'habito - - - - -	26

A similar project is underway on Comédie Française registers

See

<https://www.cfregisters.org/>



### Registres

Accédez aux versions numérisées des registres journaliers de la Comédie.



### Base de données

Accédez aux données qui concernent les recettes journalières de la Comédie.

ANNEE 1772 à 1773. DÉPENSE JOURNALIÈRE. N°. 1. De Samedi 17. avril 1772. A Représentations pour l'ouverture d'Alzire, Tragédie de M. de Voltaire et Les Femmes de bien.

	liv.	s.
A la Garde Militaire, trente-trois livres dix sols. . .	33	10
30 Jetons d'Assemblée de ce jour, à 6	180	
Jetons de lecture, à		
Jetons de comités, à		
24 Soldats assistants à chacun 1 <sup>re</sup>	24	
A. M.		
16 Feux d'Acteurs Et d'Actrices, à 20 s.	32	
Mus.		
Mlles.		
Douville		
Le Kain		
Balthazar		
pruvile		
Bryant		
Cholot		
Daubouval		
Salanval		
Mouvet		
Balthazard		
Bonnet		
Total . . .	269	10

Attesté par nous Semainiers, la dépense de ce jour d'un vingt-huit  
avril mil sept. cent. soixante-trois montant à la somme  
de Deux cent. sixante-neuf livres dix sols.

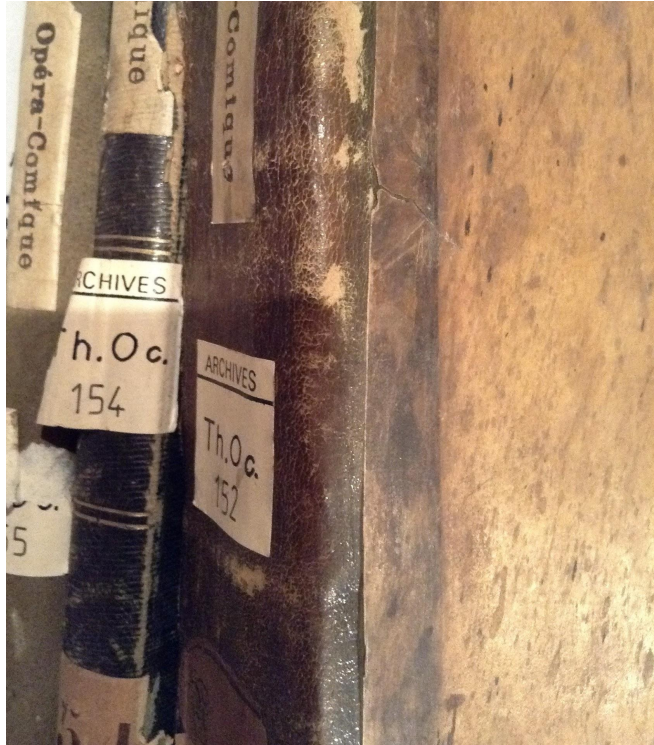
Dubouval Salanval

# The corpus

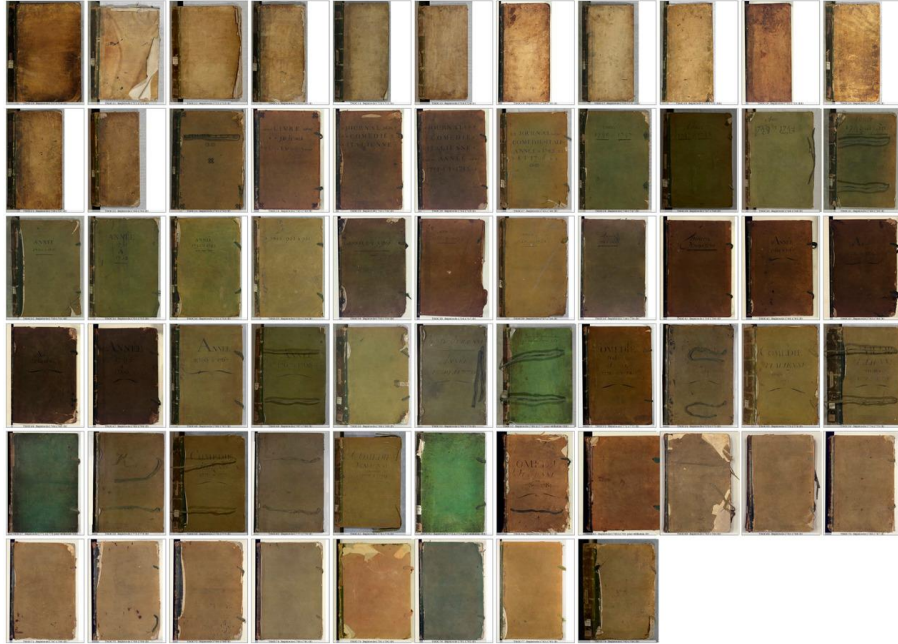


# Accounting registers

kept under the roof of the  
Bibliothèque-musée de l'Opéra

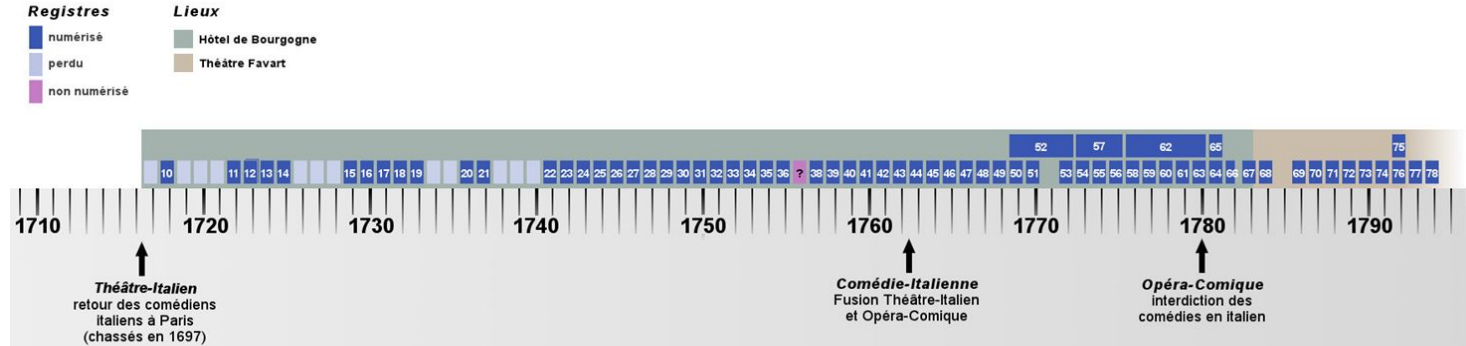


# Available corpus



Registers for Comédie-Italienne - now available on [Gallica](#)

# A century of accounting registers



(figure by Florent Coubard)

- 1 Register = 1 Theatre Season, from April to March
- digitized corpus: 1717-1794 (with 13 seasons missing)
- 64 registers of 300 to 600 pages
- about 26 000 pages

# Anatomy of a register

*Le Dimanche 4 May 160*  
*Les Depenses en monnaie*  
*compte de l'année avec le Comptable*

Date	Description	Montant
10. Mai	...	...
11. Mai	...	...
12. Mai	...	...
13. Mai	...	...
14. Mai	...	...
15. Mai	...	...
16. Mai	...	...
17. Mai	...	...
18. Mai	...	...
19. Mai	...	...
20. Mai	...	...
21. Mai	...	...
22. Mai	...	...
23. Mai	...	...
24. Mai	...	...
25. Mai	...	...
26. Mai	...	...
27. Mai	...	...
28. Mai	...	...
29. Mai	...	...
30. Mai	...	...
31. Mai	...	...

Daily accounts  
77%

*Compte General de May 160*  
*Recettes*  
*Depenses*

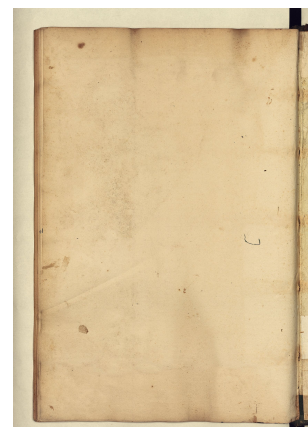
Date	Description	Montant
1. Mai	...	...
2. Mai	...	...
3. Mai	...	...
4. Mai	...	...
5. Mai	...	...
6. Mai	...	...
7. Mai	...	...
8. Mai	...	...
9. Mai	...	...
10. Mai	...	...
11. Mai	...	...
12. Mai	...	...
13. Mai	...	...
14. Mai	...	...
15. Mai	...	...
16. Mai	...	...
17. Mai	...	...
18. Mai	...	...
19. Mai	...	...
20. Mai	...	...
21. Mai	...	...
22. Mai	...	...
23. Mai	...	...
24. Mai	...	...
25. Mai	...	...
26. Mai	...	...
27. Mai	...	...
28. Mai	...	...
29. Mai	...	...
30. Mai	...	...
31. Mai	...	...

Monthly accounts  
10%

*Le Roy*  
*Recettes*  
*Depenses*

Date	Description	Montant
1. Mai	...	...
2. Mai	...	...
3. Mai	...	...
4. Mai	...	...
5. Mai	...	...
6. Mai	...	...
7. Mai	...	...
8. Mai	...	...
9. Mai	...	...
10. Mai	...	...
11. Mai	...	...
12. Mai	...	...
13. Mai	...	...
14. Mai	...	...
15. Mai	...	...
16. Mai	...	...
17. Mai	...	...
18. Mai	...	...
19. Mai	...	...
20. Mai	...	...
21. Mai	...	...
22. Mai	...	...
23. Mai	...	...
24. Mai	...	...
25. Mai	...	...
26. Mai	...	...
27. Mai	...	...
28. Mai	...	...
29. Mai	...	...
30. Mai	...	...
31. Mai	...	...

Final state + various  
3%



Blank pages  
10%

TH-OC-42 register (1760-1761) - 373 pages

# Original corpus characteristics

- Handwritten pages
- 2+ languages: French and Italian (various dialects)
- 7+ « bookkeepers »: from Alborghetti to Linguet
- Currency of the Ancien Régime : Livre-Sou-Denier
- All along the century
  - formal changes of balance sheets
  - updating of accounting rules



# Digitizing approaches

Two complementary approaches

Automated (AI-based)

vs

Manual (expert+crowdsourcing)

# AI automation

Usage of AI for segmentation and transcription

- PhD work by Adeline Granet during the project (2015-2018)
- Difficult corpus characteristics + lack of ground truth (bootstrap)

Off-the-shelf open solutions now exist for HCR

- **Transkribus**: mostly open-source, but recent paying model for using models in SAAS mode.
- **eScriptorium** (web interface for **Kraken**): open-source

# Crowdsourcing approach

- more tedious, workforce-demanding
- fit for irregular contents
- more apt at identifying exceptional/interesting items
- a requirement for building the ground-truth
- issues of data quality evaluation and data validation



# Crowdsourcing platform - ScribeAPI

ScribeAPI software

issued from the Zooniverse project

MIT-licensed

Ruby-on-Rails platform

# Crowdsourcing tasks in short



3 activities

- Mark (segment and categorize)
- Transcribe
- Verify

# Crowdsourcing interface

8 different page  
types

Déterminez le type de la page ci-  
contre :

COUVERTURE

INTRODUCTION

ETAT

COMPTE QUOTIDIEN

COMPTE MENSUEL

COMPTE ANNUEL

PAGE VIERGE

OK

Inclassable ?

# Crowdsourcing interface - marking

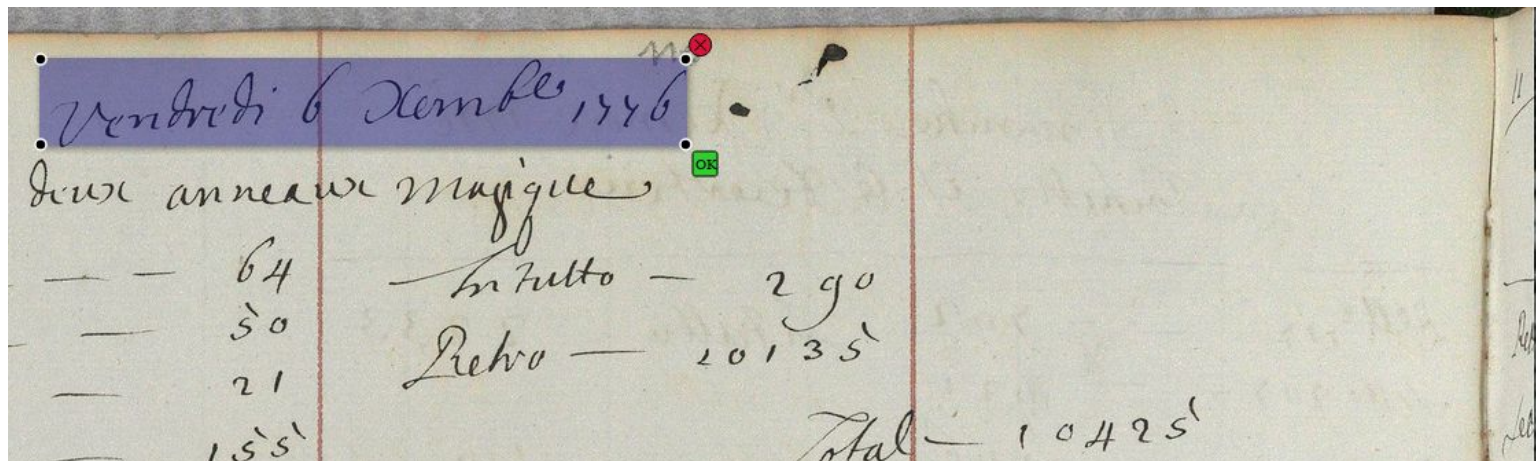
133 data categories - displayed according to the page type

Marquez les postes de recette (2/5):

- ☒ Loge particulière ?
- ☐ Théâtre
- ☐ Première loge (1)
- ☐ Seconde loge (1)
- ☐ Troisième loge (1)
- ☐ Quatrième loge
- ☐ Parterre (1)
- ☐ Supplément
- ☐ Autre recette
- ☐ Total des recettes (1)

← SUIVANT

# Crowdsourcing interface - marking

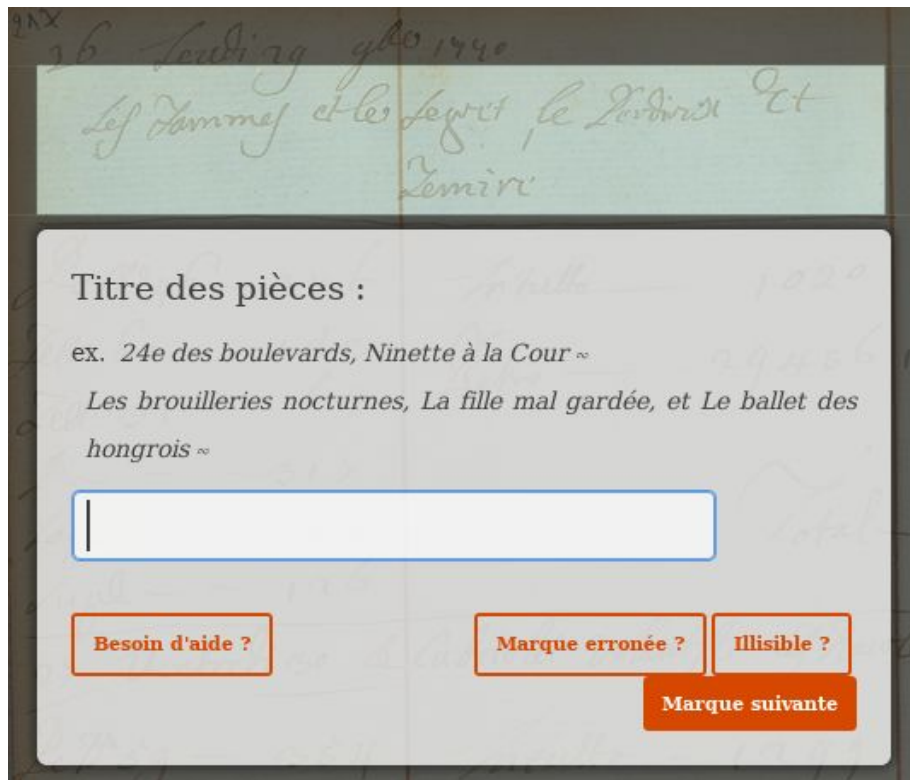


Marquez les éléments d'en-tête (1/5):

- ☐ Date du jour ?
- ☐ Titre(s)
- ☐ Événement exceptionnel
- ☐ Numéro de page

SUIVANT

# Crowdsourcing interface - transcription



The interface displays a handwritten document snippet at the top. Below it, a form for transcription is shown. The form includes a label 'Titre des pièces :', an example text 'ex. 24e des boulevards, Ninette à la Cour ~', and a list of titles 'Les brouilleries nocturnes, La fille mal gardée, et Le ballet des hongrois ~'. A text input field is provided for transcription. At the bottom, there are four buttons: 'Besoin d'aide ?', 'Marque erronée ?', 'Illisible ?', and 'Marque suivante'.

26 Lundi 29 août 1790

Les femmes et le secret, le perdrix et Zémire

Titre des pièces :

ex. 24e des boulevards, Ninette à la Cour ~

Les brouilleries nocturnes, La fille mal gardée, et Le ballet des hongrois ~

Besoin d'aide ?

Marque erronée ?

Illisible ?

Marque suivante

*Les femmes et le secret, la perdrix et Zémire*

# Crowdsourcing interface - verification



The image shows a crowdsourcing interface for verifying handwritten text. At the top, a snippet of a handwritten document is displayed, featuring the date "Du Dimanche 21 Juillet 1776" and the title "Le Roi le fermier La Parodie". Below this, an orange banner prompts users to "Votez pour la meilleure transcription". Underneath, the label "Saisie d'origine : Titre des pièces :" is followed by two selectable options: "Le roi et le fermier, La parodie" and "Le roi, le fermier, la parodie". At the bottom, two buttons are provided for further action: "Autre proposition ?" and "Marque erronée ?".

Votez pour la meilleure transcription

Saisie d'origine : *Titre des pièces :*

Le roi et le fermier, La parodie

Le roi, le fermier, la parodie

Autre proposition ?

Marque erronée ?



# ScribeAPI

## Pros

- Open source
- Collaborative annotation of images
- Able to express chained task sequences (mark-transcribe-verify)
- Well thought interface (accessible to non-technical users)

## Cons

- Complex architecture/model
- End-of-life reached in 2016



# Open-Source and Technical debt

Official ScribeAPI - End-of-Life in 2016

Dependencies on old versions of Ruby, js/coffeescript, Mongo, unbuildable as-is now

Current server/dependencies dockerized

But not all hope is lost

effort by Utrecht University to update critical dependencies (unmerged pull request)

# After crowdsourcing - ongoing work

- Data quality evaluation
  - first phase crowdsourced (peer-reviewed verification)
  - second phase automated
  - third phase manual
- Data cleaning and validation
  - custom tools, using external information (list of play titles, actors...)
  - dedicated user-interface (dashboard) for experts
  - integration of AI-based tools
- Data publication
  - FAIR principles
  - collaboration/links with other projects/data

# If you want to contribute, welcome at...

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

