RECITAL
Combining crowdsourcing and expertise in Digital Humanities

FOSDEM - February 2021

Olivier Aubert
www.olivieraubert.net - 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
CETHEFI - Centre d’Études 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)
A similar project is underway on Comédie Française registers

See https://www.cfregisters.org/
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
- 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

Daily accounts 77%
Monthly accounts 10%
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.
- **eScripotorium** (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
Crowdsourcing interface - marking

133 data categories - displayed according to the page type
Crowdsourcing interface - marking
Les femmes et le secret, la perdrix et Zémire
Crowdsourcing interface - verification

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


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/