



Decision Support in Crowdsourcing

The RECITAL Workshop

Olivier Aubert³, Benjamin Hervy¹, Guillaume Raschia³ and Françoise Rubellin²

contact : guillaume.raschia@ls2n.fr

24 sept. 2021

1. Centre François Viète (CFV) - EA CNRS 1161; Université de Nantes, France
2. L'Antique et le Moderne (LAmO) - EA CNRS 4276; Université de Nantes, France
3. Laboratoire des Sciences du Numérique de Nantes (LS2N) - UMR CNRS 6004; Polytech Nantes, France

{C}ontrainte et {I}ntégration :
pour une {RE}évaluation des {S}pectacles {F}orains et {I}taliens

French ANR Contract 2015-2018

- Study of fair and Italian theaters
 - Restricted theater
 - Contrivences and subterfuges to get around (banners, pantomime)
- **Quantitative History**
 - Economy of live performance
 - From large and original corpus
- Focus on
 - **Registers of la Comédie-Italienne in the 18th century**

{RE}gisters of la {C}omédie-{ITAL}ienne



- Primary source, preserved under the roof of Musée de l'Opéra (BnF), Paris
- Also available online at gallica.bnf.fr
- Each register = 1 Theater Season, from **April to March**
- Balance sheet by daily show
- Time range : **1717 to 1794**

Anatomy of a Register

25 250 pages in 64 volumes (~400 pages each)



● daily budget

● monthly budget

● yearly budget

● blank

● misc

TH-OC-14

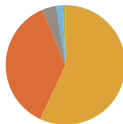
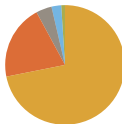
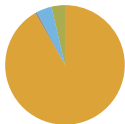
TH-OC-26

TH-OC-45

TH-OC-55

TH-OC-69

TH-OC-71



Crowdsourcing platform RECITAL

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



Features

- Powered by [ScribeAPI](#) - the early [Zooniverse](#) technology
- 3 activities : **Mark**, **Transcribe**, **Verify**
- 8 categories of documents, **134** labels for annotations

Le Mardi 5^e fevrier 1754 *Ceul page*

quatorz^e debut de M^{lle} Catinore
La Mere confidente, L'amant auteurs

	Mespr. ———	2 ^{te}
	feux ———	29 10
	Domestiques ———	12
	Elleues ———	
	Revierquier ———	3
	Payé ala Garde ———	33 10
	acteur qui ont joué	
17 Premieres ———	68	
15 Secondes ———	30	

Marquez les éléments d'en-tête (1/5):

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

SUIVANT

[Consulter un tutoriel](#)

☒ Transcrire cette page

☐ À propos de ce registre

Rechercher une discussion

[Forum](#)

Une dépense : [Quantité :]

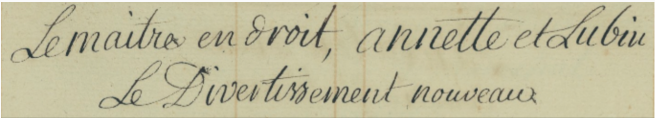
Libellé :

Montant :
ex. 32 ~ 32:8:2 ~ 32::2 ~ ::2 ~

Besoin d'aide ? Marque erronée ? Illisible ? Marque suivante

Trompette
à mannes
à m. Carot
Premier
Grandes
Supplément

3
4 10
7 i
12 5 7 11
60
30
67
12
159



Le maître en droit, annette et Lubin
Le Divertissement nouveau

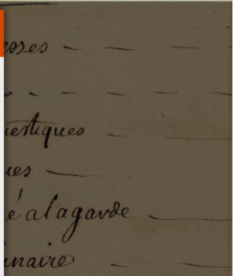
Votez pour la meilleure transcription

Saisie d'origine : Titre des pièces :

Le maître en droit, annette et Lubin Le Divertissement nouveau

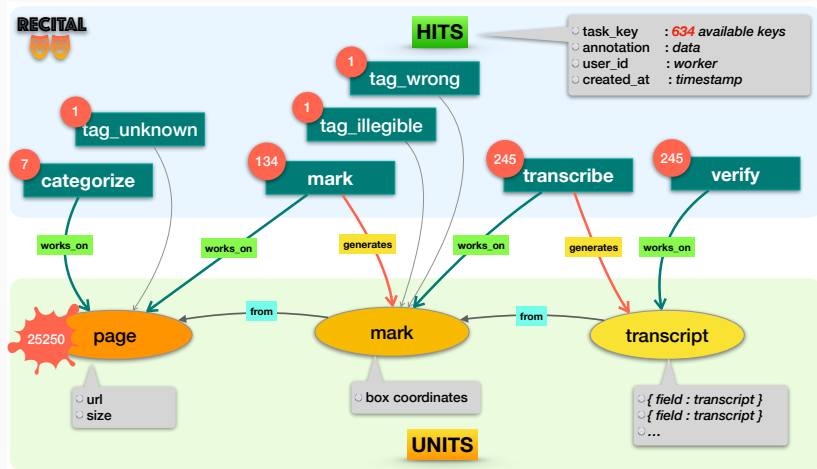
le maître en droit, Annette et Lubin, le divertissement nouveau

Autre proposition ? Marque erronée ?



roses -
estiques -
es -
e' alagarde -
inaire -

Overview of the ScribeAPI Data Model



[2021-09-09] RECITAL has more than 2 500 workers, 950K hits and 640K units

Data Quality Assessment, Data Cleaning

Topics : Database, Analytics, NLP, ML

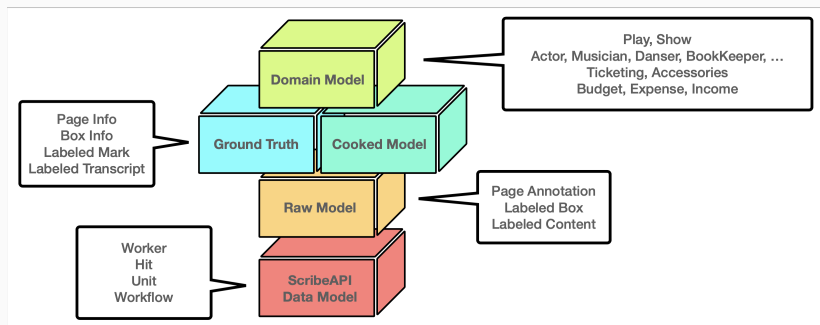
1. **Basic Text Cleaning** (conservative/lossy process), then
2. **Inter-Annotator Agreement**. On failure, then
3. **Conflict Resolution** (aggressive/lossy process, fuzzy matching)
4. **Record Linkage** to map same entities (plays, actors, ...)

Requirements

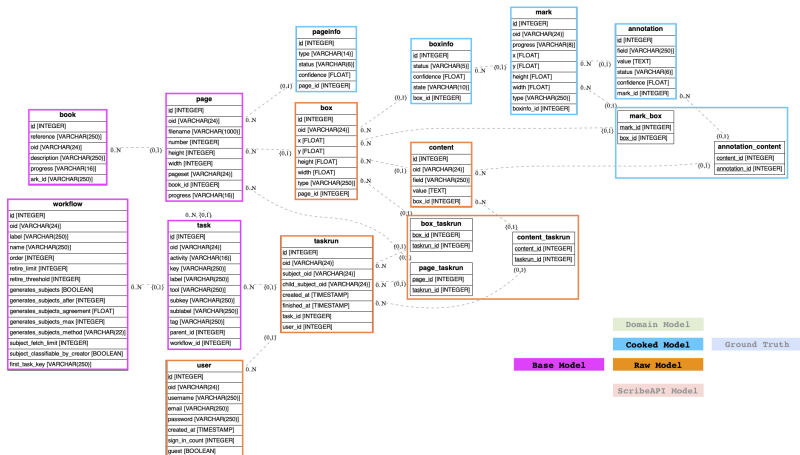
1. Design an effective data model to preserve the data-to-document link;
2. Perform a 3-layers **Extract-Transform-Load** process up to the « Domain Model ».
3. Develop monitoring tools

The RECITAL Data Model

From the ScribeAPI Crowdsourcing Data Model to Plays and Shows!



The RECITAL Logical Model for Real



Base Model defines root units (books-pages) and CS settings (workflows-tasks)

The RECITAL Data Model - Excerpt

950K hits

1-500 of 954 489							
ORDER BY							
id	oid	subject_oid	child_subject_oid	created_at	finished_at	task_id	user_id
415	589b87d277616c3856740100	589a5db677616c37fb366500	589b87d277616c3856770100	2017-02-08 21:04:18.734000	2017-02-08 21:04:19.337000	22	8
416	589b87db77616c3856780100	589a5db677616c37fb366500	589b87db77616c3856790100	2017-02-08 21:04:27.162000	2017-02-08 21:04:27.778000	23	8
417	589b87e477616c38524f0100	589a5db677616c37fb366500	589b87e477616c3852500100	2017-02-08 21:04:36.683000	2017-02-08 21:04:37.250000	25	8
418	589b87f477616c3852510100	589a5db677616c37fb366500	589b87f477616c3852520100	2017-02-08 21:04:52.510000	2017-02-08 21:04:53.075000	26	8
419	589b87f477616c3852530100	589a5db677616c37fb366500	589b87f477616c3852540100	2017-02-08 21:05:01.342000	2017-02-08 21:05:01.954000	27	8

430K marks

1-500 of 427 051							
ORDER BY							
id	oid	x	y	height	width	type	page_id
1	589af2c377616c3856150000	927.635359116022	2169.742868564735	94.93861229810332	290.7513812154697	db_total_revenues	10018
2	589af3c077616c38529a0000	1286.36344617058913	1186.2503425171735	135.45922561434773	1076.1322010431522	db_guard	10018
3	589af3fd77616c38530e0000	1041.81379231878956	748.9835991522931	119.75683996957592	736.62768608779935	db_fire	13870
4	589b02577616c3851110000	104.0086278300725	669.83552168291403	387.7238381133666	7668.561109576566	db_fire	14674

285K transcripts

1-500 of 285 686							
ORDER BY							
id	oid	field	value				
1	589af3a777616c3856170000	db_total_revenues	414				1
2	589af3d677616c38520c0000	db_guard_amount	33:10:				2
3	589b0dce77616c3852130000	db_title	Ninette à la Cour, L'école des mères, et Les chasseurs avec le ba...				4
4	589b0dd777616c38561a0000	db_fire	37:10				3
5	589af3d677616c38520c0000	db_guard_amount	33:10				2
6	589h11a677616c38524d0000	db_month	août				5

recital.univ-nantes.fr/dashboard

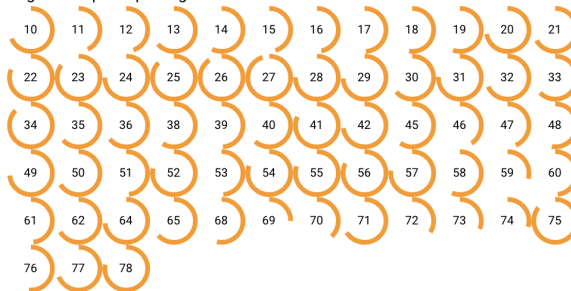
Focus on the Cooked Model

Avancement du marquage par registre

Nombre moyen de pages marquées sur l' **ensemble des registres**



Pages marquées par registre



RECITAL Dashboard - Register View

571 pages dont 518 pages marquées.

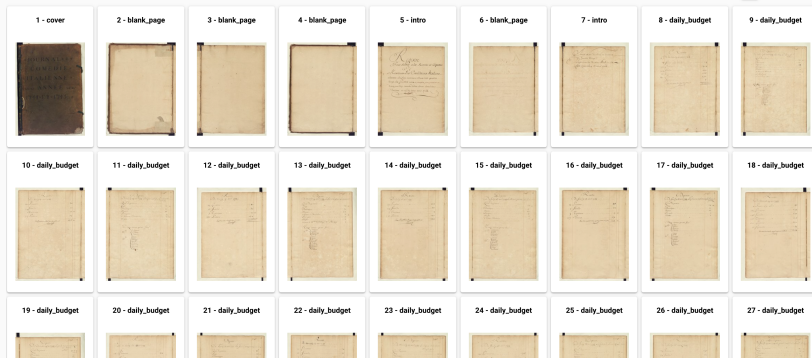
PAGES MARQUÉES



VUE GRILLE



Search

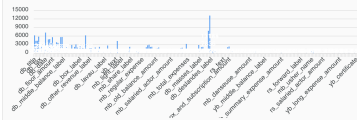


RECITAL Dashboard - Transcript View

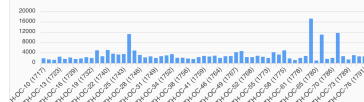
228540 contenus		Champ	Valeur		
Id	Action	Champ	Valeur	Registr	Page
23443	P E R	db_title	fortunate grelozie	10	thoc10-005
23450	P E R	db_third_box_num	1	10	thoc10-005
23451	P E R	db_third_box_amc	1:10	10	thoc10-005
23455	P E R	db_total_revenues	387:10	10	thoc10-005
23456	P E R	db_tax	96:17.6	10	thoc10-005
125728	P E R	db_mances_amo	2	10	thoc10-007
125729	P E R	db_mances_qty	4	10	thoc10-007
125730	P E R	db_mances_label	mances	10	thoc10-007
58802	P E R	db_date	lundi 7 juin 1717	10	thoc10-009
58803	P E R	db_title	li due anelli magici	10	thoc10-009
58804	P E R	db_total_expenses	3:122:1	10	thoc10-009
58837	P E R	db_first_box_num	29	10	thoc10-009
58838	P E R	db_first_box_amc	58	10	thoc10-009
59455	P E R	db_floor_num	74	10	thoc10-009

228540 / 228540 contenus

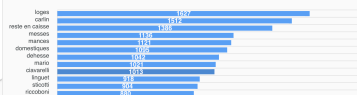
Nombre de marques par champ



Nombre de marques par registre

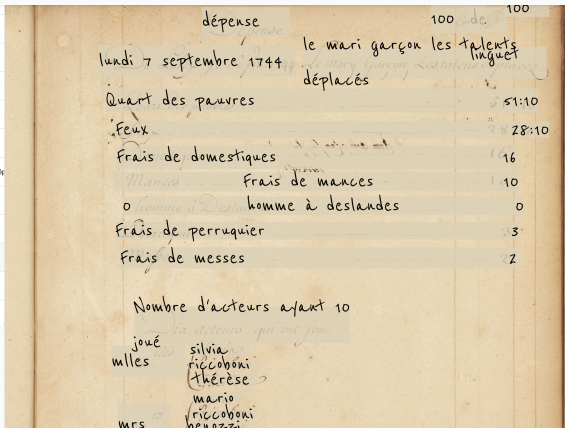


Termes (filtrés) récurrents



Back to sources

☐	1681x78+493+961	db_deslandes	0 homme à deslandes 0
☐	1696x79+455+859	db_mances	10
☐	1676x97+468+748	db_domestic	16
☐	1817x73+463+645	db_fire	28:10
☐	1814x101+422+509	db_tax	51:10
☐	1021x132+1249+336	db_title	le mari garçon les talents dé
☐	823x132+395+328	db_date	lundi 7 septembre 1744
☐	368x124+826+168	db_text	dépense
☐	157x106+1948+168	db_cashier	paraphe de linguet
☐	228x121+1783+156	db_page_number	100
☐	182x81+2099+130	db_page_number	100



What we have technically achieved?

- data : two SQL databases (main and ground truth)
- post-proc : A bunch of Python scripts
- devops : an automated process to perform the all ETL chain at once
- web backend : a data API to access the SQL DB
- web frontend : a dashboard to explore, filter (and edit) the DB

What remains to be done?

- tons of log records from the ETL to scan
- improve on the validate/fix feature
- finish the crowdsourcing job!
- boost with HTR (eScriptorium, Transkribus)
- develop domain model analytics

Thanks!

`guillaume.raschia@ls2n.fr`

Populate the Raw Model

- Collect and organize entries of the **(840K)** worker's activity log
- Collect units **(630K)**

Under the Hood

- Split into **950K fine-grained tasks** : quantity, label, amount
- Slightly twist the model from **Mark-Transcribe-Verify** to **Categorize-Mark-Tag-Transcribe**
- Detect (and drop) scribeAPI **anomalies**
 - dangling hits : no unit created
 - missing data : box coordinates, content value, label, page category
 - missing meta-data : worker id, timestamp, workflow

Populate the Cooked Model

The cleaning step due to Crowdsourcing

- Clean page tasks → page tag or wrong or unknown
- Clean box tasks → box coordinate+label and safe or duplicate or wrong or illegible
- Clean content tasks (hard part) → transcript

Tasks are dependent from page up to transcript : preserve that chain

- discard marks from wrong pages, and so on...

Keep track of every processing. Evaluate **confidence**

Patch the Cooked Model with Ground Truth

- A side database to record each and every expert validation/correction.
- The all cooked DB may be regenerated from scratch, still **preserving expert work**
- Need
 1. id mapping (using persistent mongo object id's from scribeAPI) and
 2. a merging process to incorporate the ground truth to the cooked DB.

The Domain Model

Populate plays, actresses/actors, shows, budgets (income and expense).

- Figure out the all timeline : assign Show to Date
- Assign Tx to Budget, deal with monthly Tx
- Assign Budget to Show
- Map Show to Plays and People (Actors, Dansers, Musicians, Ushers, ...)

Keep track of every processing. Evaluate **confidence**

Preserve the all chain from the document to the domain entity