



# **Training Tesseract OCR Models for Latvian Fraktur Scripts Via Crowdsourcing**

---

valdis.saulespurens@lnb.lv

R&D at National Library of Latvia

ELAG 2022 conference



# OCR improvement pipeline

- Ingest new data with old model
- Web Application [frakturs.lnb.lv](http://frakturs.lnb.lv) to curate new data
- Train Tesseract with newly labeled data
- Repeat until needed accuracy





# History of Fraktur in Latvia

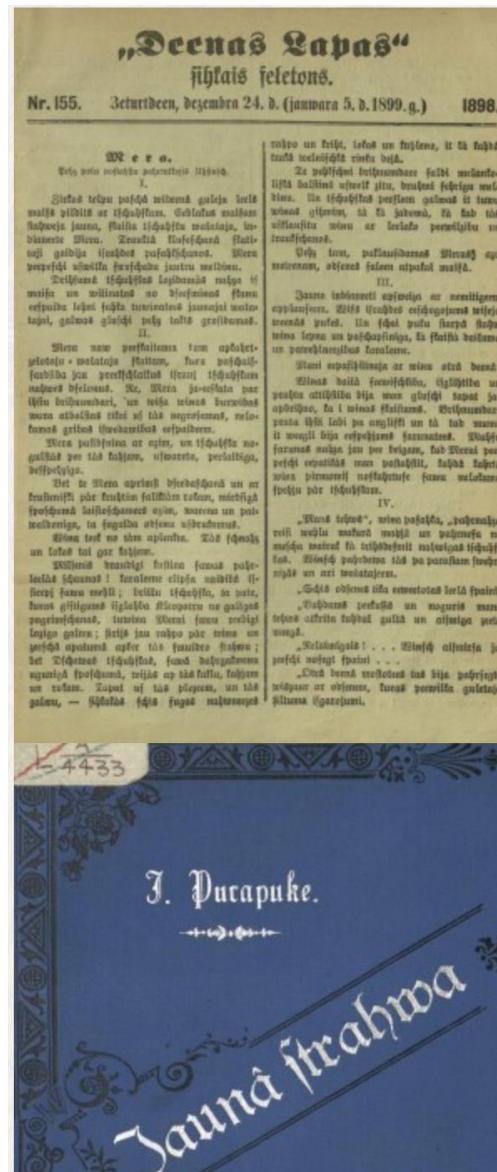
Adopted from German Fraktur 17-18th century

Contains extra Latvian only glyphs


Long s with stroke S - U+A7A8, f - U+1E9C

Standard until 1908

Still used until 1938







# Collection of Latvian Fraktur books and periodicals at NLL

---

- Conservative estimates
- 180+ books 19th-20th century
- 24 000+ periodicals



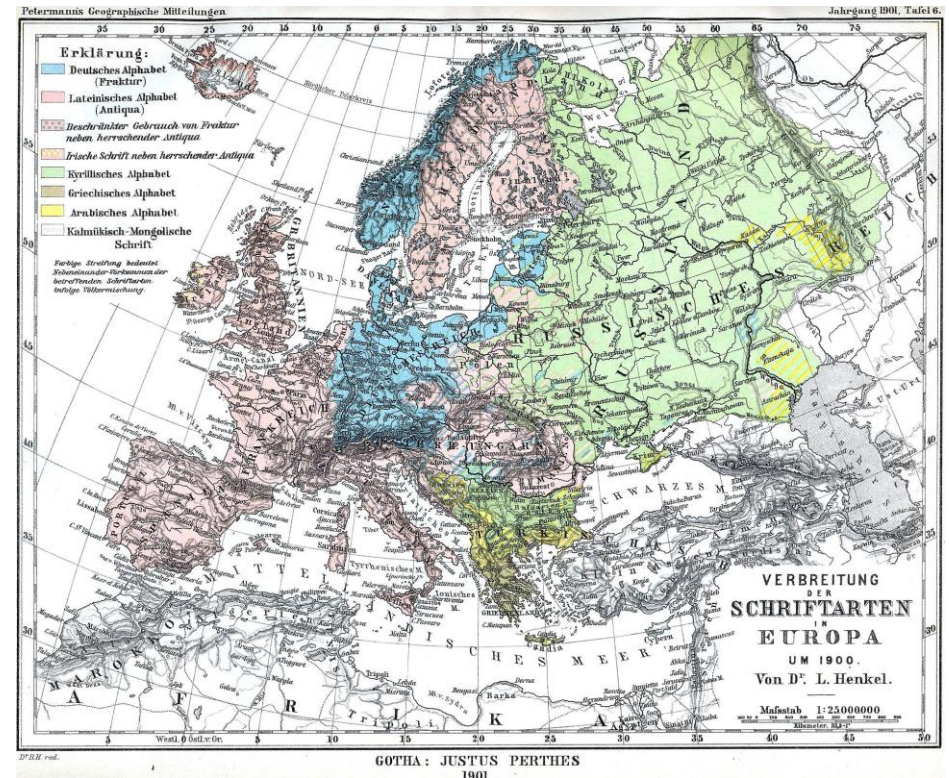
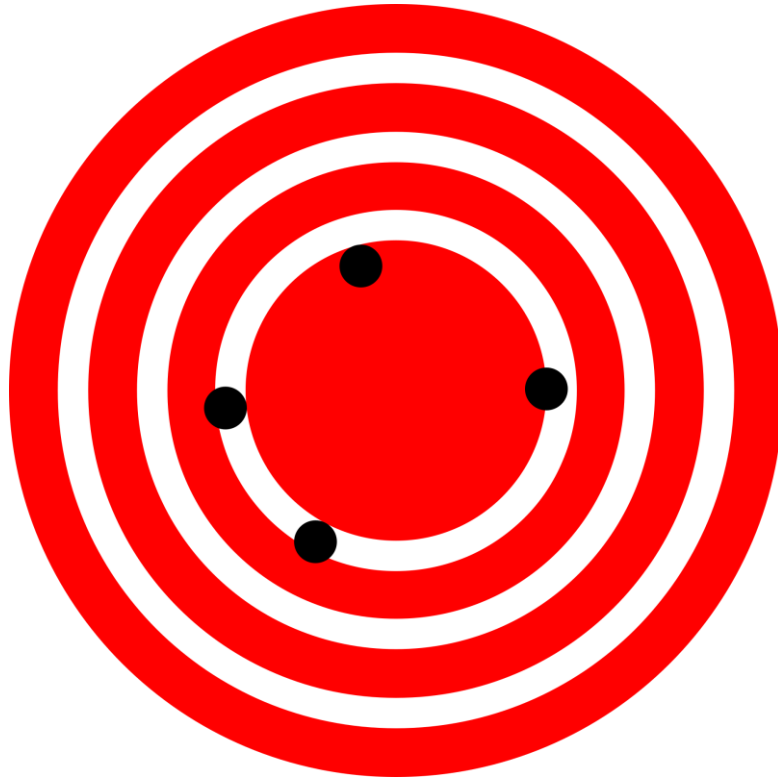


# Previous attempts at Fraktur OCR

Circa 2010 – commercial offering

Low accuracy – around 90%

Special symbols a problem



# Tesseract improvements

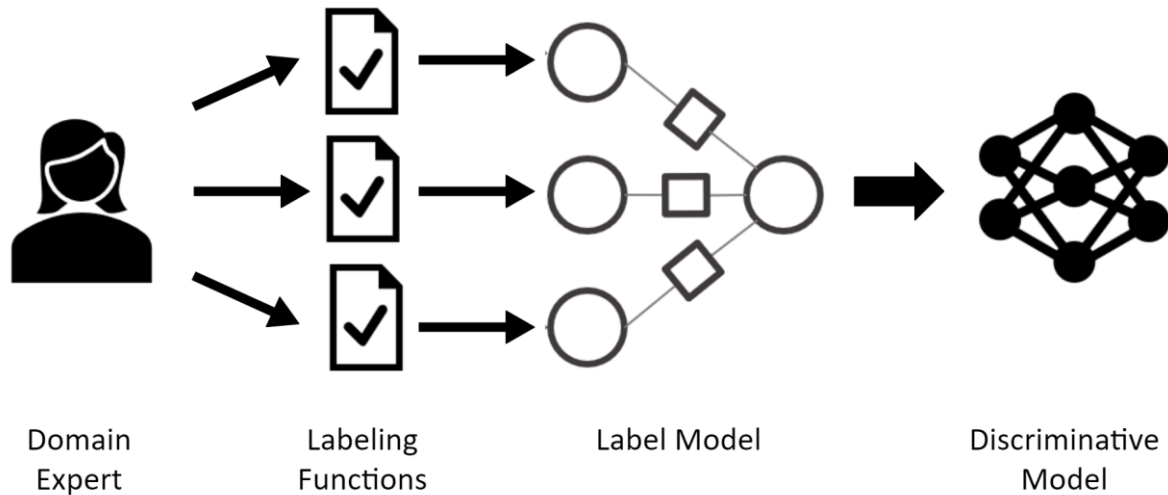


UNIVERSITÄTSBIBLIOTHEK  
MANNHEIM

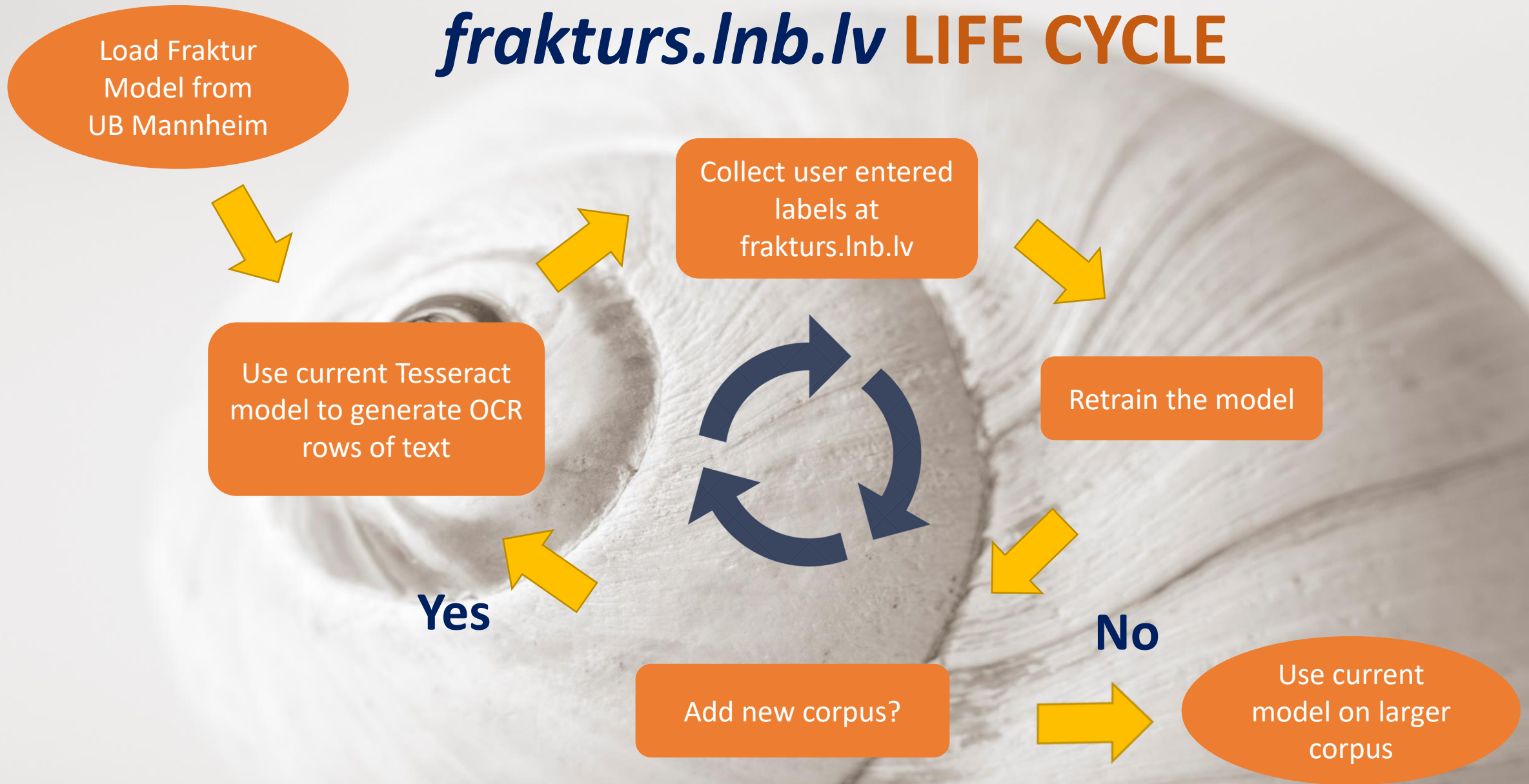
versions 4.0+  
LSTM support  
Accuracy gains

Models ->

# Need for labeling



# *frakturs.Inb.lv* LIFE CYCLE

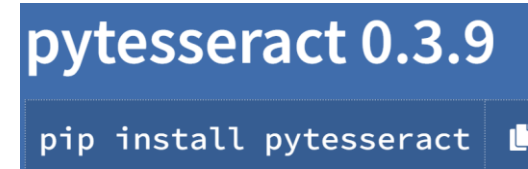




# Pre-processing

---

- Python script
- OCR on folders
- Each page extracted as dictionary
- Cut with Pillow
- Text saved into SQLite database



# Web app stack

---

---

**Use**      Use boring (proven) technology –  
boringtechnology.club

---

**Back  
End**      Python  
Flask/SQLite/Plotly/Pandas/Pillow/pytesseract

---

**Front  
End**      vanilla Javascript/jQuery/Bootstrap/Plotly

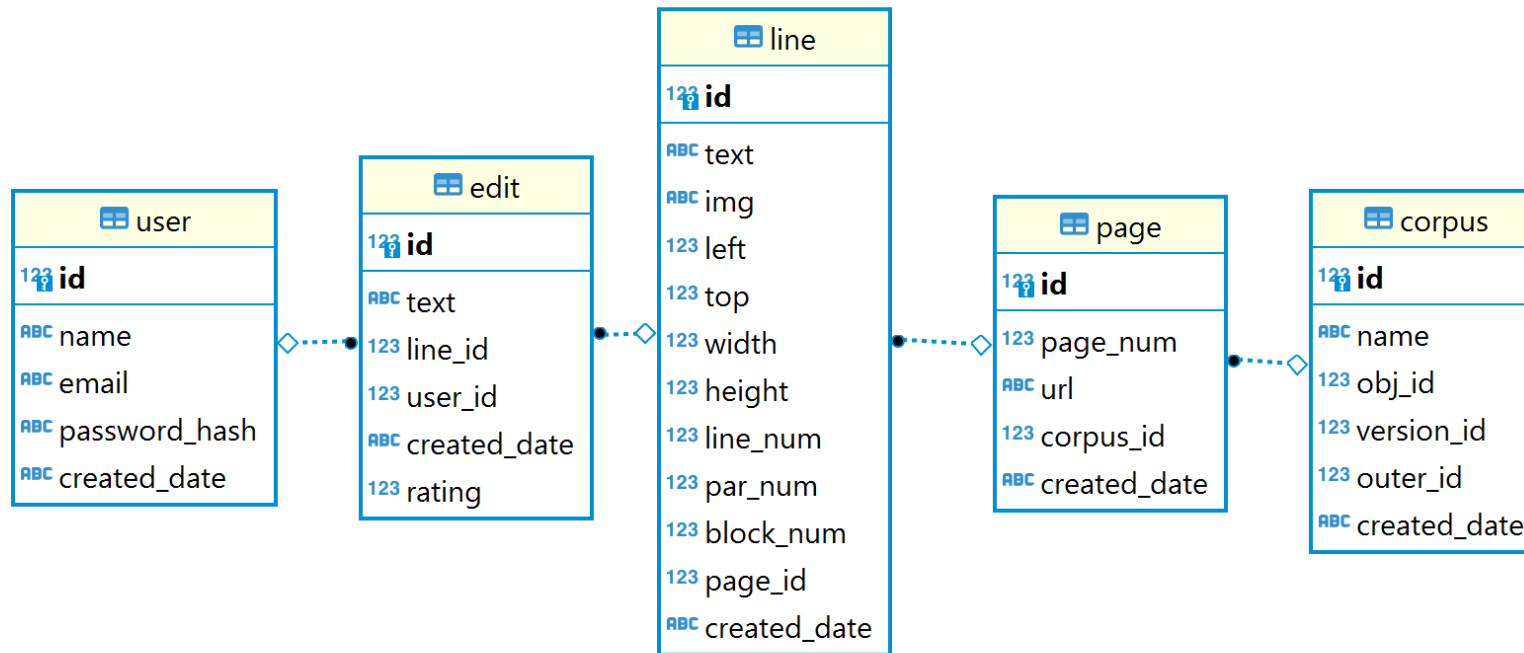
---





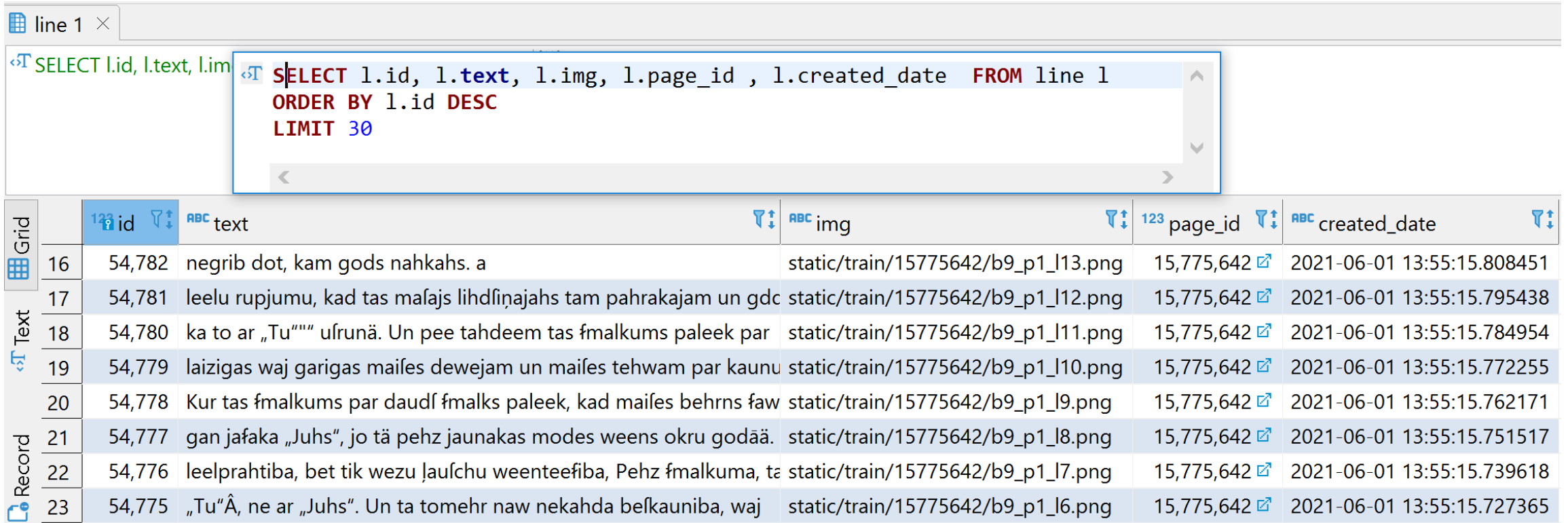
# SQLite

## Database Schema



# Rows of Text

- Each row – base for edit
- Text from previous model



The screenshot shows a database interface with a SQL query editor and a data grid. The query editor displays the following SQL query:

```
SELECT l.id, l.text, l.img, l.page_id, l.created_date FROM line 1
ORDER BY l.id DESC
LIMIT 30
```

The data grid below the query editor shows the results of the query. The columns are: id, text, img, page\_id, and created\_date. The rows are numbered 16 through 23.

	id	text	img	page_id	created_date
16	54,782	negrib dot, kam gods nahkaks. a	static/train/15775642/b9_p1_l13.png	15,775,642	2021-06-01 13:55:15.808451
17	54,781	leelu rupjumu, kad tas mafajs lihdñajahs tam pahrakajam un gdc	static/train/15775642/b9_p1_l12.png	15,775,642	2021-06-01 13:55:15.795438
18	54,780	ka to ar „Tu” ufrunä. Un pee tahdeem tas fmalkums paleek par	static/train/15775642/b9_p1_l11.png	15,775,642	2021-06-01 13:55:15.784954
19	54,779	laizigas waj garigas maifes dewejam un maifes tehvam par kaunu	static/train/15775642/b9_p1_l10.png	15,775,642	2021-06-01 13:55:15.772255
20	54,778	Kur tas fmalkums par daudf fmalks paleek, kad maifes behrns faw	static/train/15775642/b9_p1_l9.png	15,775,642	2021-06-01 13:55:15.762171
21	54,777	gan jafaka „Juhs”, jo tä pehz jaunakas modes weens okru godää.	static/train/15775642/b9_p1_l8.png	15,775,642	2021-06-01 13:55:15.751517
22	54,776	leelprahtiba, bet tik wezu ļaufchu weenteefibā, Pehz fmalkuma, ta	static/train/15775642/b9_p1_l7.png	15,775,642	2021-06-01 13:55:15.739618
23	54,775	„Tu” ne ar „Juhs”. Un ta tomehr naw nekahda beļkauniba, waj	static/train/15775642/b9_p1_l6.png	15,775,642	2021-06-01 13:55:15.727365



# SQLite Views

---

gd_curators
123 id
ABC text
123 line_id
123 user_id
ABC created_date
123 rating

gd_edits
123 id
ABC text
123 line_id
123 user_id
ABC created_date
123 rating

gd_ratings
123 id
ABC text
123 line_id
123 user_id
ABC created_date
123 rating

mostly_perfect_lin...
123 id
ABC text
123 line_id
123 user_id
ABC created_date
123 rating

mostly_extra
123 id
ABC text
123 line_id
123 user_id
ABC created_date
123 rating

perfect_match
123 id
ABC text
123 line_id
123 user_id
ABC created_date
123 rating

match_curat...
123 id
ABC text
123 line_id
123 user_id
ABC created_date
123 rating

match_text
123 id
ABC text
123 line_id
123 user_id
ABC created_date
123 rating

## Front End

- Simple HTML templates
- Javascript/jQuery
- Bootstrap for CSS – no design budget for this
- Thumbs Up / Down



fkolotajs un ķesteris. Leels labdaris nelaikis bijis fkolai. Elfa  
ifolotajs un Ffeferis. Ceels labdaris nelaikis bijis fkolai. Elfa

Pārbaudīts  Nesaprotams

Pareizs teksts: fkolotajs un ķesteris. Leels labdaris nelaikis bijis fkolai. Elfa

\s S

]s f

[s f

Saglabāt

Izlabotas rindīņas: 143132



# Javascript with jQuery

- Keyboard processing of rare symbols
- Define new jQuery methods

gihmi un raustofchos apakfchas luhpu, zelahs gan laikam no preefch-  
gihmi un raustofchos apakfchas luhpu, zelahs gan laikam no pi

  Pārbaudīts   Nesaprotams

pusdeenu darbofchanahm, fludinafchanahm jeb gara fatrizinafchanas,  
pusdeenu darbofchanahm, fludinafchanahm jeb gara fatrizinafch

  Pārbaudīts   Nesaprotams

kas beigās padara nokufumu. Pehzpusdeenās un naktīs meeru bau-  
kas beigās padara nokufumu. Pehzpusdeenās un naktīs meeru l

  Pārbaudīts   Nesaprotams

dijufchas flimneezes gars rihtōs ir ņpirgtaks, darbigaks. Ja tahdu  
dijufchas flimneezes gars rihtōs ir ņpirgtaks, darbigaks. Ja tahdu

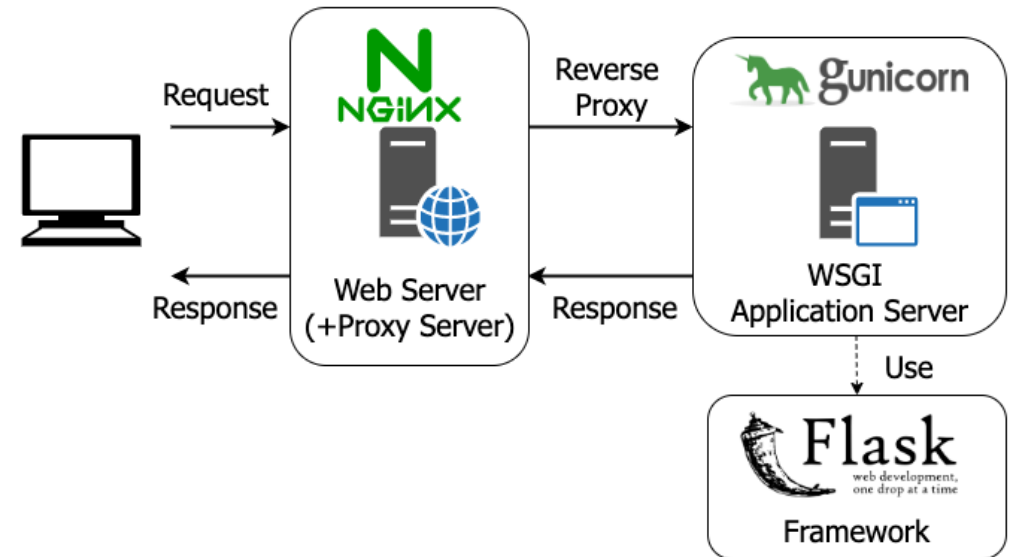
  Pārbaudīts   Nesaprotams

\s <b>S</b>	]s <b>f</b>	[s <b>f</b>		
[a <b>â</b>	]a <b>à</b>	[e <b>ê</b>	]e <b>è</b>	
[i <b>î</b>	]i <b>ì</b>	[o <b>ô</b>	]o <b>ò</b>	[u <b>û</b>

**Saglabāt**

# Deployment

National Library of Latvia Infrastructure VPS - Ubuntu 18.04LTS





# Post-processing – Fine Tuning

---

- Extract Ground Truth – 2 Matching Votes
- <https://github.com/tesseract-ocr/tesstrain/wiki/GT4HistOCR#fine-tuning-based-on-scriptfraktur>
- 50-100k iterations – ideal would be 1M or more
- Accuracy ~ 99%

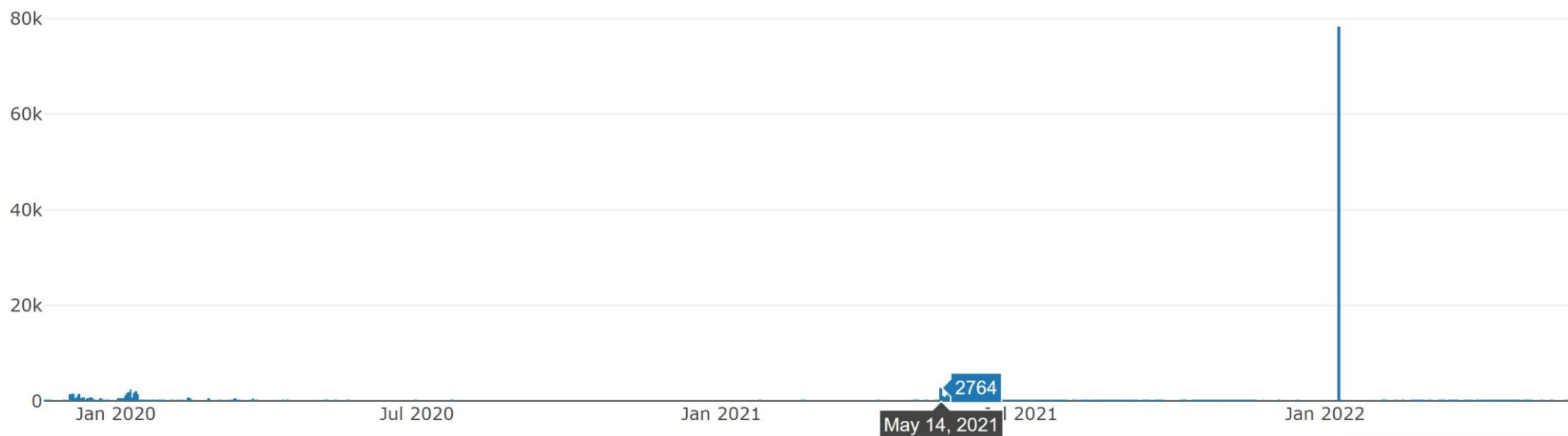


# Results

- User Activity
- Two campaigns – Thank you Anda Baklane!
- Google Bot in Jan



Labotāju aktivitāte pa dienām

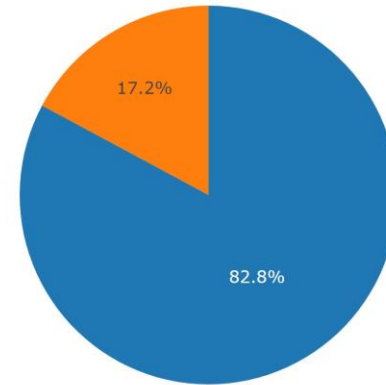
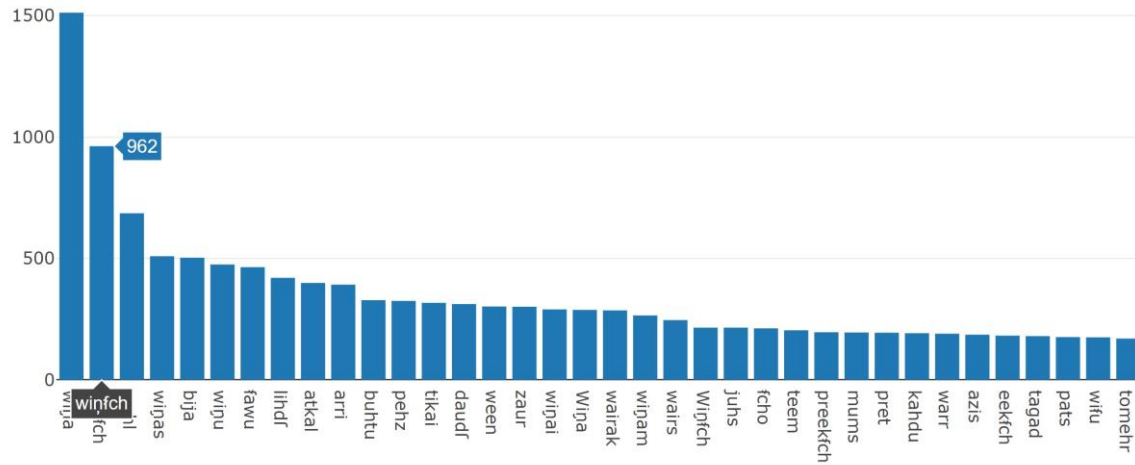




# Statistics

- Popular words
- Readability

Populārākie vārdi(>=4 un <=100 garums)



Readable  
Unreadable

# Key lessons

---

- Sense of purpose – more engagement (70+ participants)
- Pareto principle – 80% of edits by less than 20% of participants




# Future

---

- Streamline adding of new books
- Continuous Integration of new models
- Solve double symbol issue
- Upgrade of Tesseract engine





Thank You !  
Questions and  
comments

---

Aaâà	Aaâà	Mm	Mm
Bb	Bb	Nn	Nn
Cc	Cc	Nn	Nn̄
Dd	Dd	Doõ	Ooô
Eeêè	Eeêè	Pp	Pp
Ff	Ff	Rr	Rr
Gg	Gg	Rr	Rr̄
Gg	Gġ	Ss	Ss
Hh	Hh	f	f
Iiû	Iiû	Sf	Sf
Jj	Jj	Tt	Tt
Kk	Kk	Uuû	Uuû
Kk	Kk̄	Vv	Vv
Ll	Ll	Ww	Ww
Ll	Ll̄	Zz	Zz