

# Searching Extensive Archives

with the help of **#AI** tools

27. 10. 2023

Martin Bulín  
bulinm@kky.zcu.cz

Jan Švec  
honzas@kky.zcu.cz

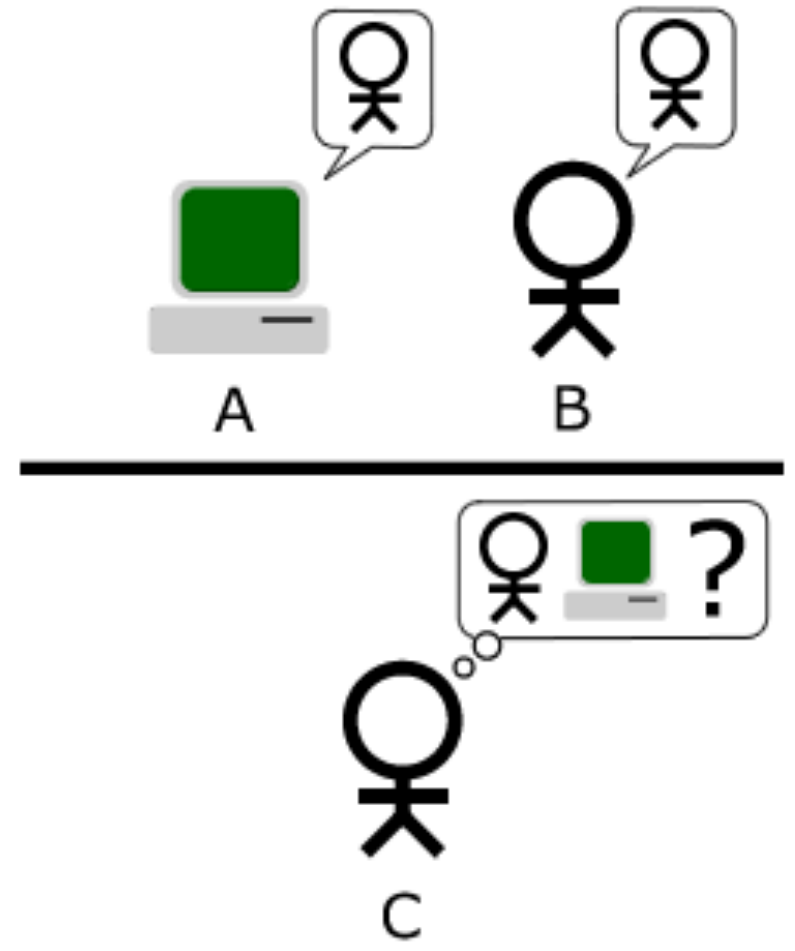
Pavel Ircing  
ircing@kky.zcu.cz

# The AI Phenomenon

As I see it...

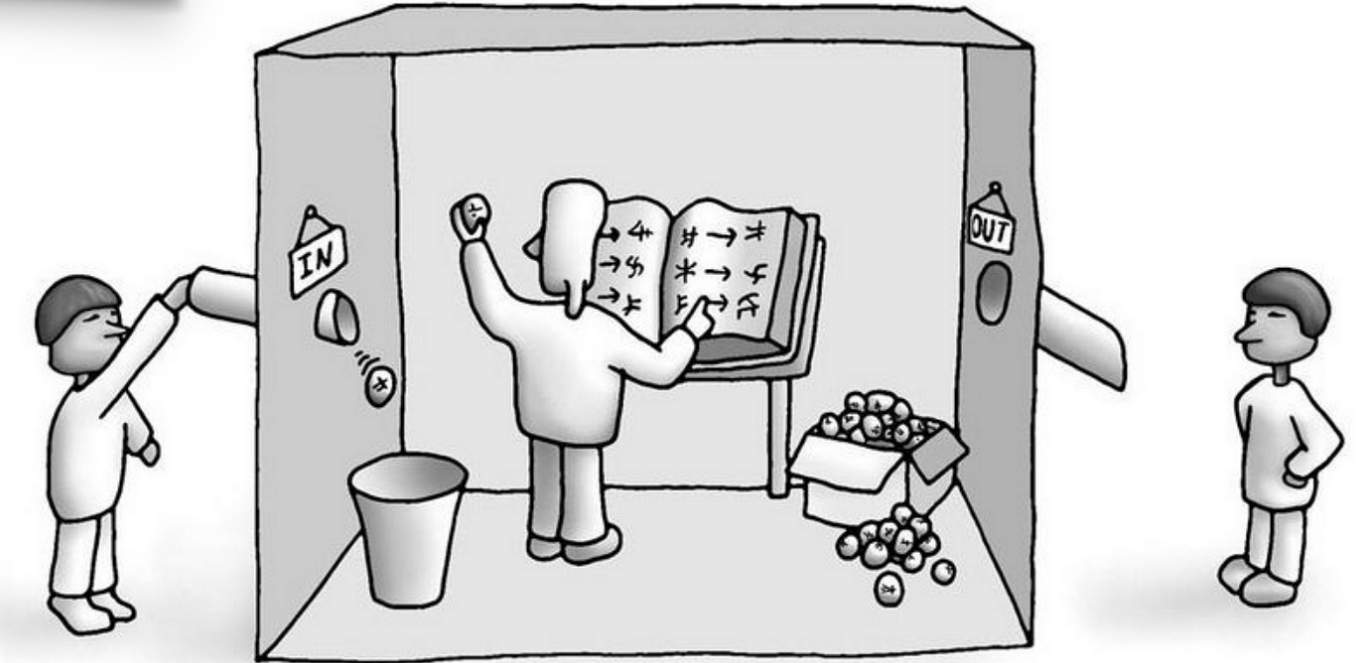
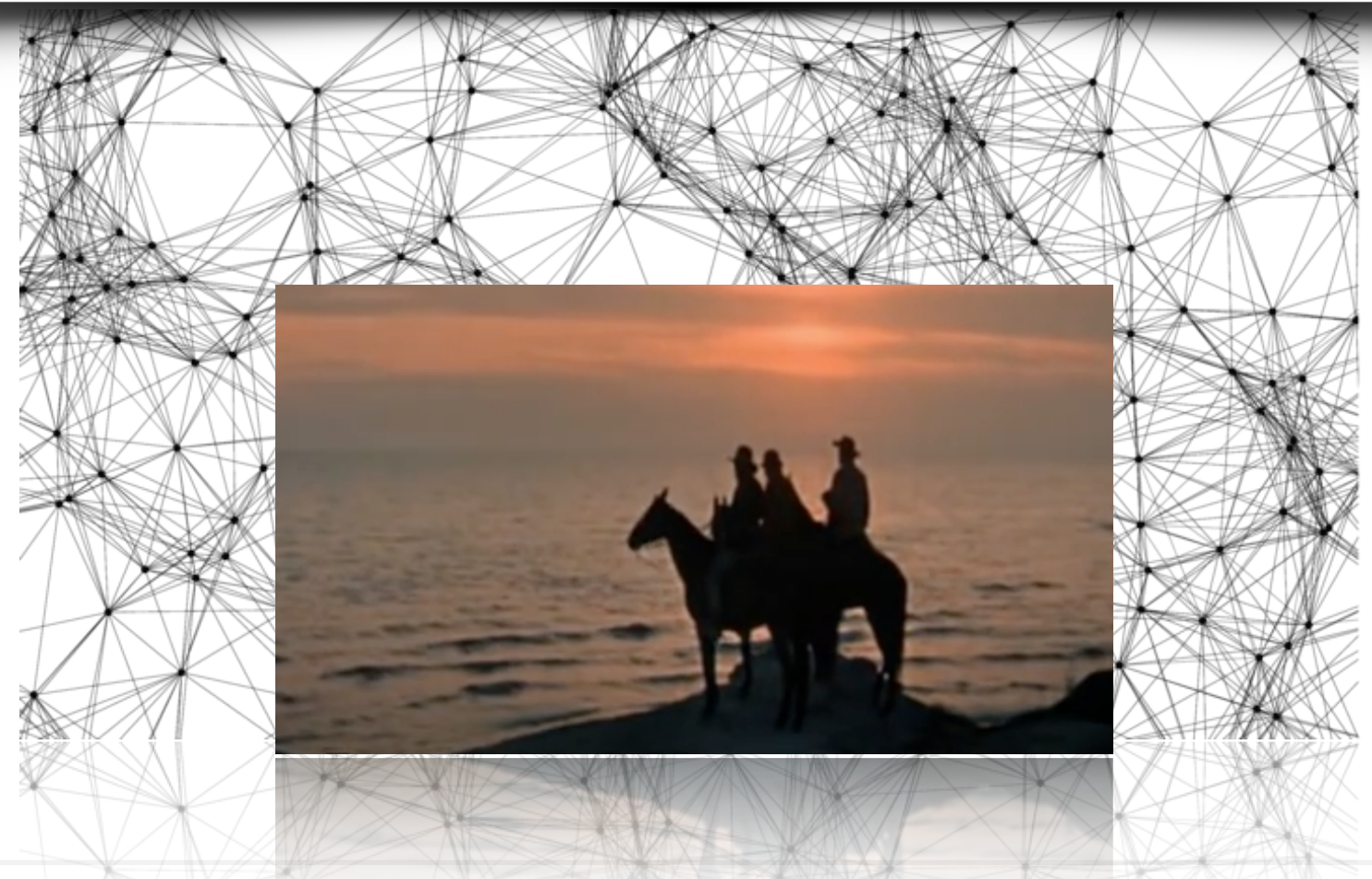


# ChatGPT



Applied  
~~Artificial~~ ~~Idea~~  
 Intelligence

- great technologies / ideas / math
- => targeted applications for **#AI tools**



# Cooperation of *KKY* with Institutions

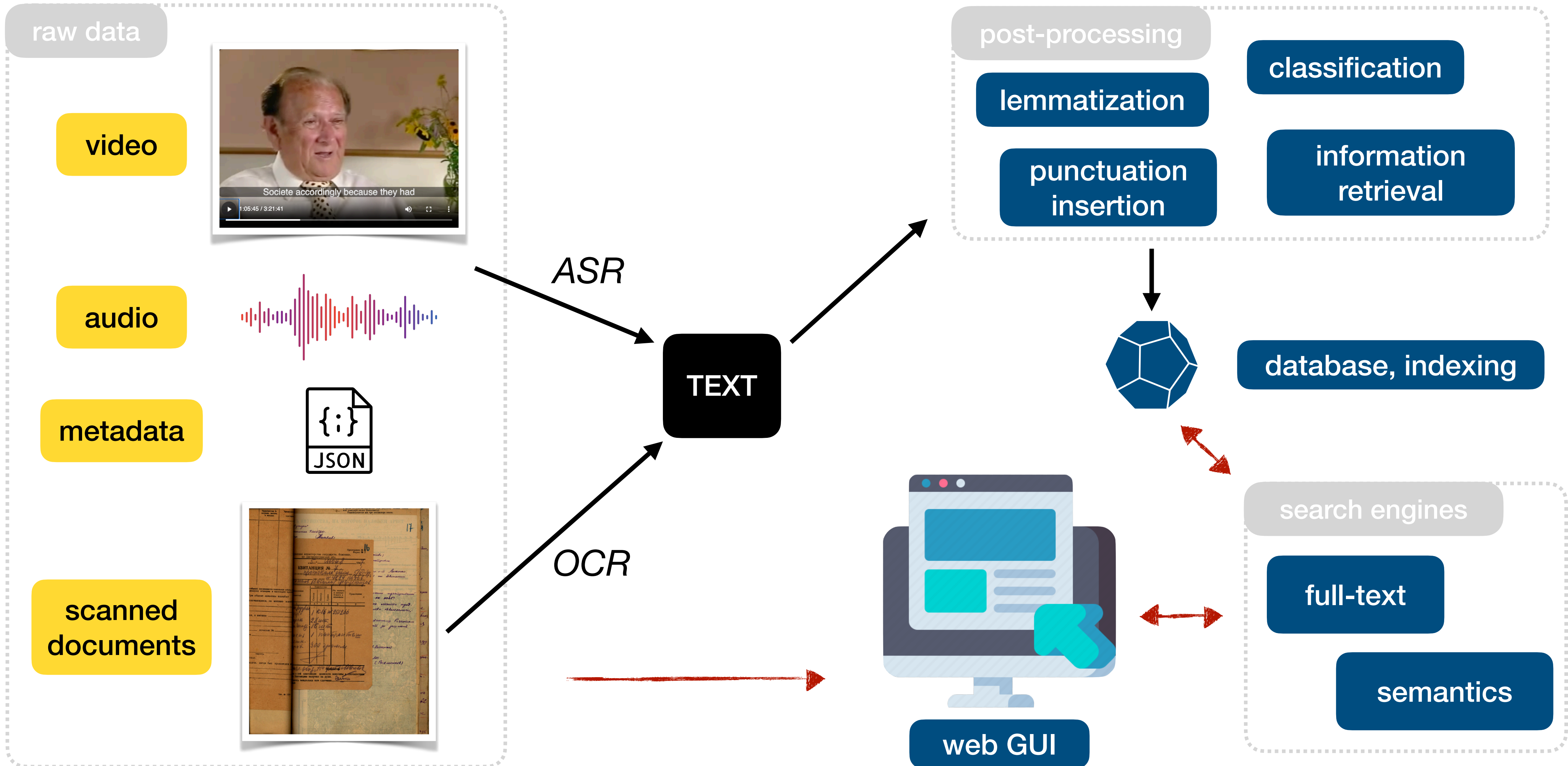
## focused on digitization and preservation of history



The **MALACH** project (NSF, 2001 - 2006)  
 Multilingual Access to Large Spoken Archives

# KKY Technologies for Working with Large Archives

## General workflow



# KKY Technologies for Working with Large Archives

## Automatic Speech Recognition (ASR)

- state-of-the-art **#AI** method (Wav2Vec 2.0) and... **"tailored" for the task!**
- stable general models for Czech, Slovak, English and German at the moment
- no hallucinations and *"faster than real-time" @ CPU*

Method name	# of params	Year	English	Czech	German	Slovak
GMM-HMM diag	7.6M	2004	39.60	40.23		
GMM-HMM full	19M	2013	34.27	26.00		
DNN-HMM	31M	2017	30.91	23.18		
TDNN-LF-MMI	8.1M	2020	20.79	17.44		
CNN-TDNN-LF-MMI	6.8M	2021	17.85	14.65		
<b>Wav2Vec 2.0</b>	<b>95M</b>	<b>2023</b>	<b>12.88</b>	<b>8.43</b>	<b>17.08</b>	<b>11.57</b>
Wav2Vec-XLS-R	300M		14.31	9.50	22.52	12.17
Whisper-large	1550M		17.34	25.95	22.99	27.49

[% WER]

results on

MALACH data

✓ API

✓ Web Interface



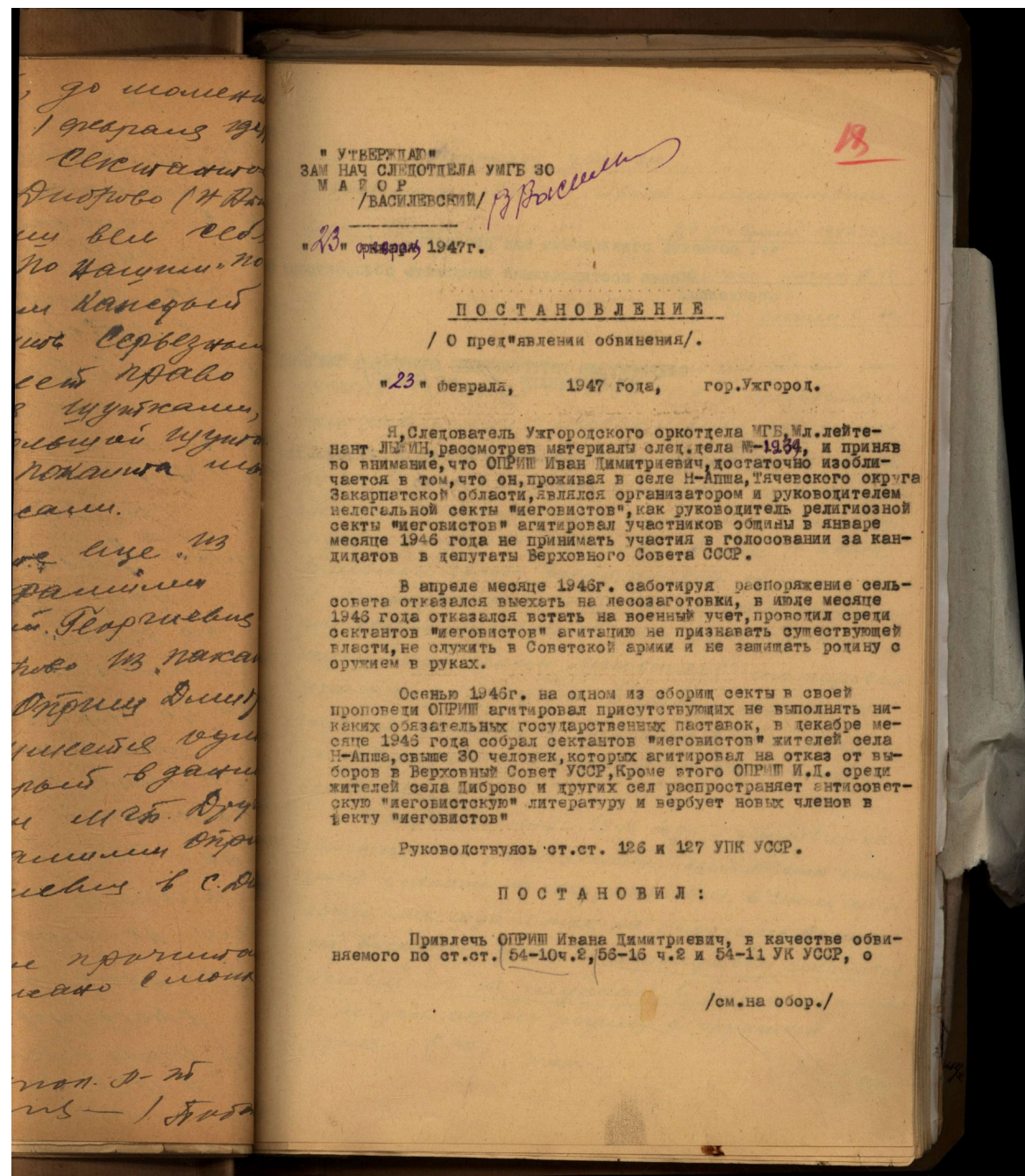
[uwebasr.zcu.cz](http://uwebasr.zcu.cz)

# KKY Technologies for Working with Large Archives

## Optical Character Recognition (OCR)

<https://tesseract-ocr.github.io/>

- application and customisation of the Tesseract library (#AI)
- two phases: multilingual model (language recognition) + language specific model
- this is actually a *bottleneck* in our pipeline (especially the handwritten OCR)



ПОСТАНОВЛЕНИЕ

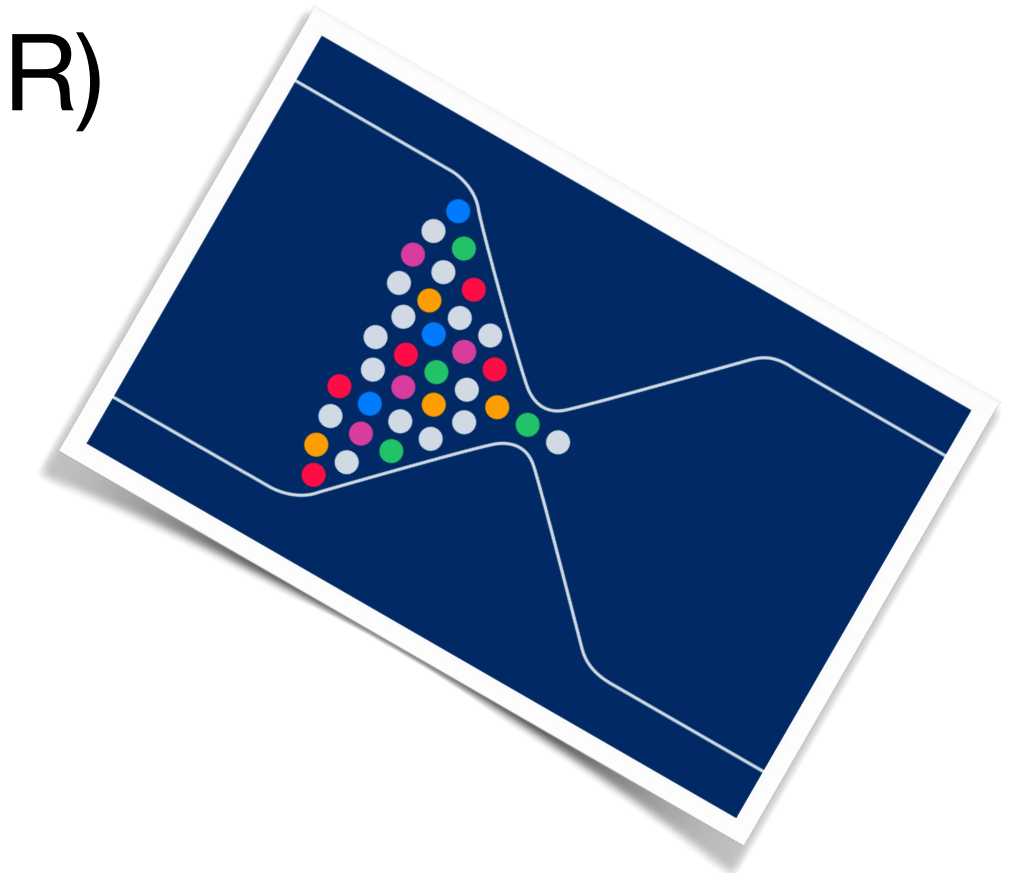
ЫЫЫ

/ О прежиявлении обвинения./

"23 февраля, 1947 гола, гор.Ужгород.

Я,Слекователь Ужгородского оркотцела МГБ Мл .лейте-  
нант ЛЫИН, рассмотрев материалы слек: дела №1994, и приняв  
во рнимание, что ОПРИ" Иван Лимитриевич, хостаточно изобли-  
чается в Том, что ОН, проживая в селе Н-Апша, Тячевского округа  
Закарпатской области, являлея организатором и руководителем  
нелегальной секты тиеговистов", как руководитель религиозной  
секты чиеговистови агатировал участников эбцины в январе  
месяце 1946 года не принимать участия в голосовании за кан-  
цидатов в цепутаты Верховного Совета СССР.

В апреле месяце 1946г. саботируя оеспоряжение сель-  
согета отказался выехать ва лесозаготовки, в июле месяце  
1946 года отказался встать на военный учет, проводил среди  
сектантов "иеговистов" агитацию не призывать сутествующей  
власти, не служить в Советской армии и ве затитать родину в  
оружием в руках.



*any ideas?  
experience?*

# KKY Technologies for Working with Large Archives

## Post-processing of the ASR (or OCR) output

★ semantic level

- BERT-based models **#AI**
- punctuation marks detection (period, comma, question mark) and capitalization
- supported languages at the moment: **cs, en, de, sk**

... jewish to their roots my father sent money regularly to what he called the chalutzim chalutzim or is a hebrew word for pioneers my grandparents were very much concerned with the pioneers in israel in fact as children we were told that the money we put into the blue box was going for purchasing land from the sultan for the jewish people in palestine ...



... Jewish to their roots. My father sent money regularly to what he called the Chalutzim chalutzim, or is a Hebrew word for pioneers. My grandparents were very much concerned with the pioneers in Israel. In fact, as children we were told that the money we put into the blue box was going for purchasing land from the sultan for the Jewish people in Palestine. ...

# KKY Technologies for Working with Large Archives

## Generation of semantically relevant questions for given contexts

- T5-based models (**#AI**)
- generated questions are accompanied with short answers

★ helps model training

context

Ed Ryder plays the trumpet. He was sentenced to Graterford Penitentiary in Pennsylvania for 20 years **for a murder it was later shown he did not commit**. He played jazz when he was in prison. He played jazz when he got out. And he says that it is a completely different experience playing jazz to inmates.



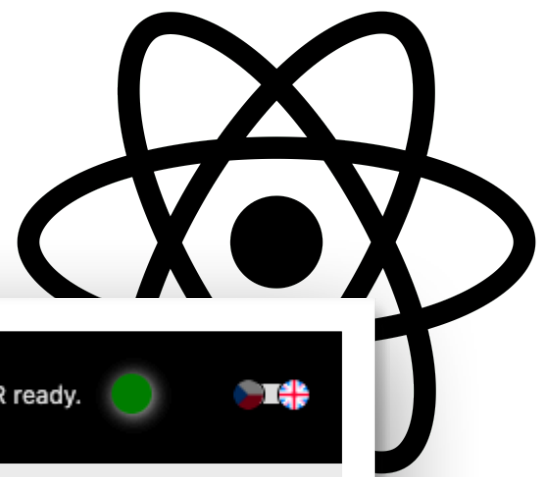
auto-generated questions

What instrument does Ed Ryder play? - Ed Ryder plays the trumpet. - How long was Ed Ryder sentenced for? - Ed Ryder was sentenced for 20 years. - Was Ed Ryder convicted of the murder he was sentenced for? - **No, it was later shown that Ed Ryder did not commit the murder** he was sentenced for.




# KKY Technologies for Working with Large Archives

## Web-based graphical user interfaces



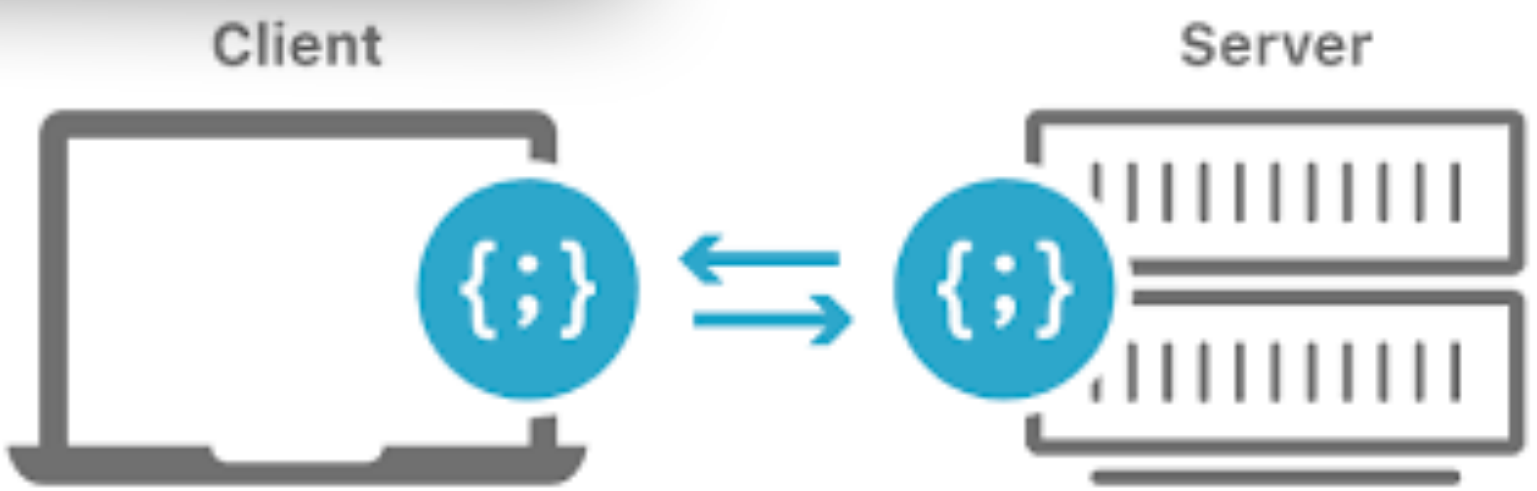
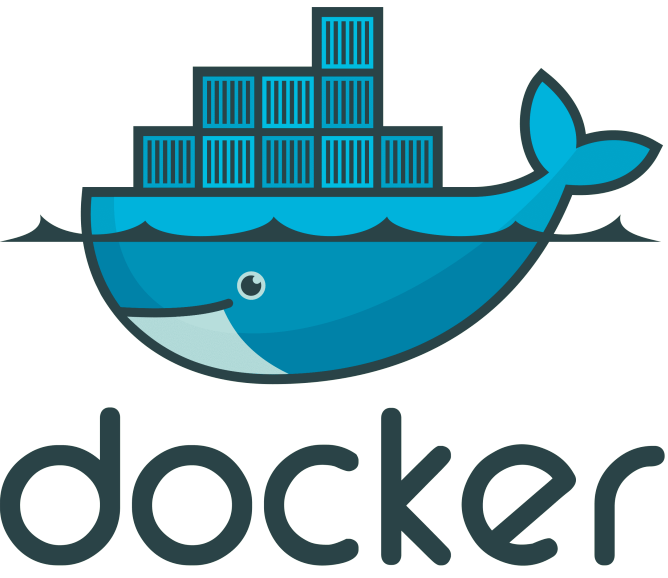
# React

DEPARTMENT OF CYBERNETICS  Semantic Search DEMO Readme Server connected and 36 interviews loaded. ASR ready.

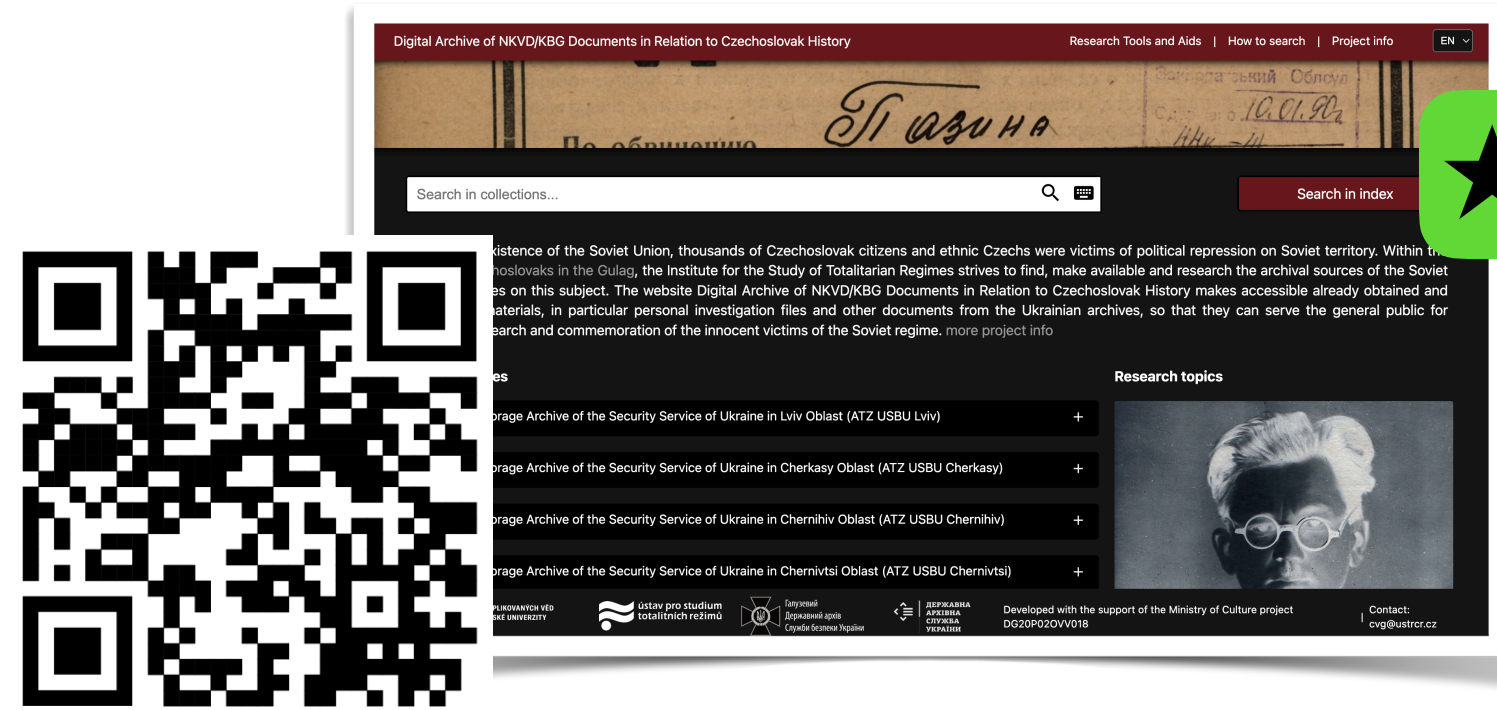
Abraham Bomba METADATA Type your query...

0:00:10	What is the country of the interviewer?	92%
0:00:39	What is the language in English?	93%
0:03:09	Did the interviewee attend a public school from a city?	99%
0:03:12	Did the interviewee attend a public school from the city?	100%
0:03:15	Did the interviewee receive a Jewish education in the sixth grade?	95%
0:04:01	Was the size of the population to the entire population significant?	

I had an attack of mine, 1:27:48 / 3:21:41

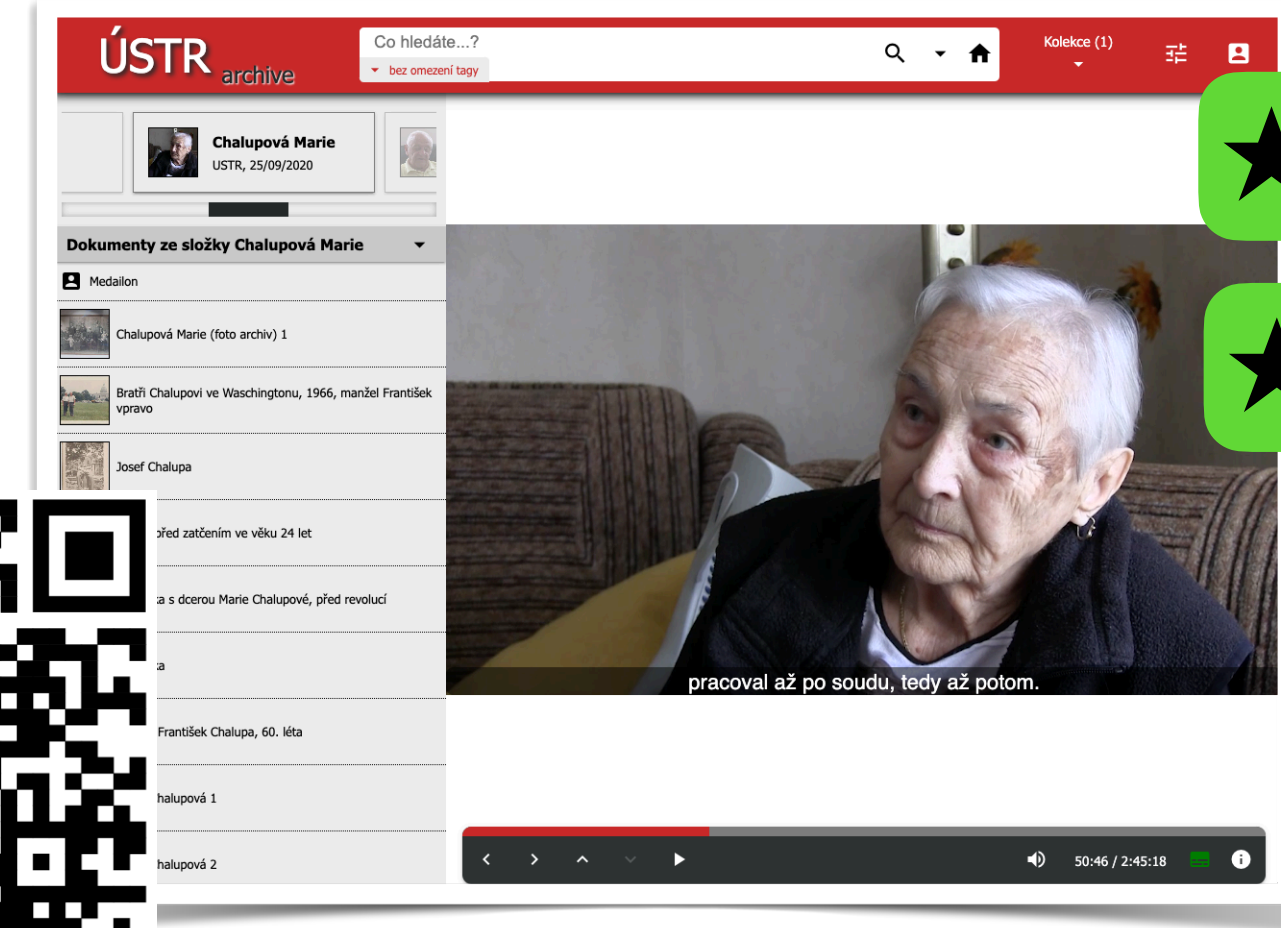


# Deployed Interfaces Run by *KKY*



★ OCR

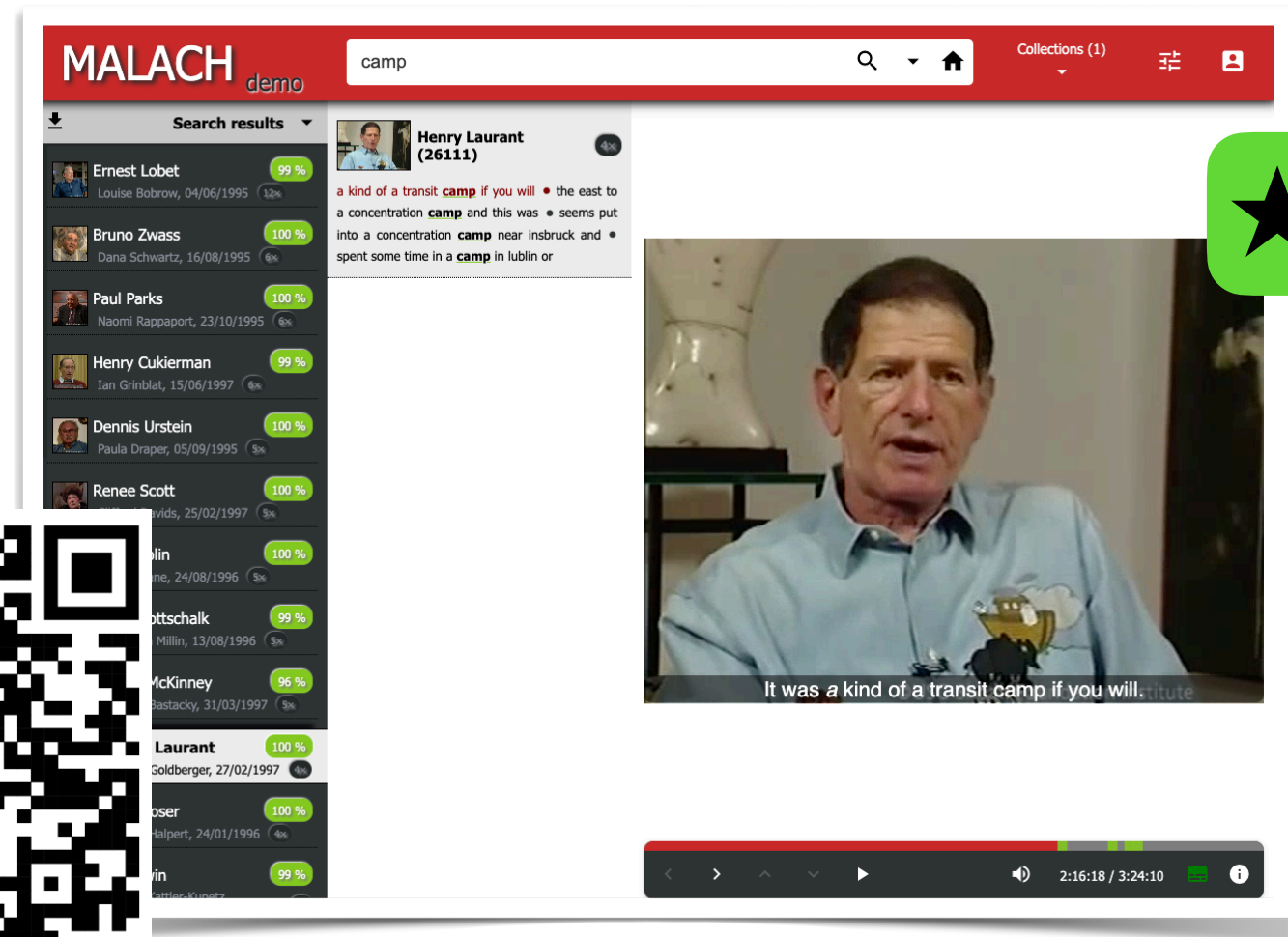
<https://archivkgb.zcu.cz>



★ ASR

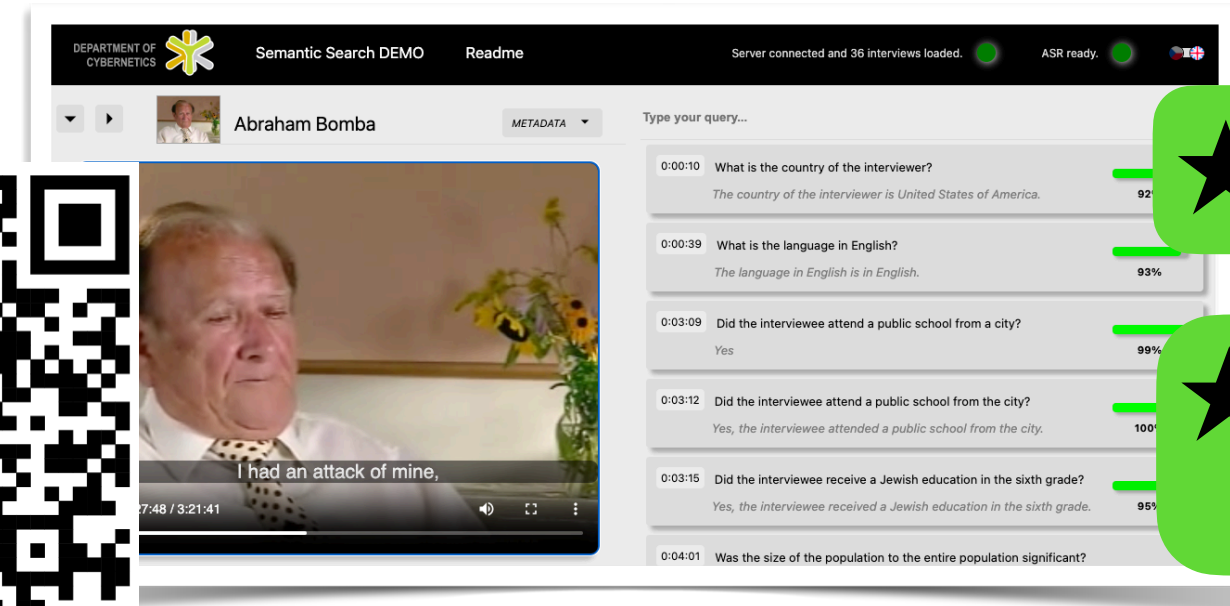
★ OCR

<https://naki-ustr.zcu.cz>



★ ASR

<https://malach.kky.zcu.cz>



★ ASR

★ Semantic Search

<https://malach-aq.kky.zcu.cz>

# Digital Archive of NKVD/KGB Documents

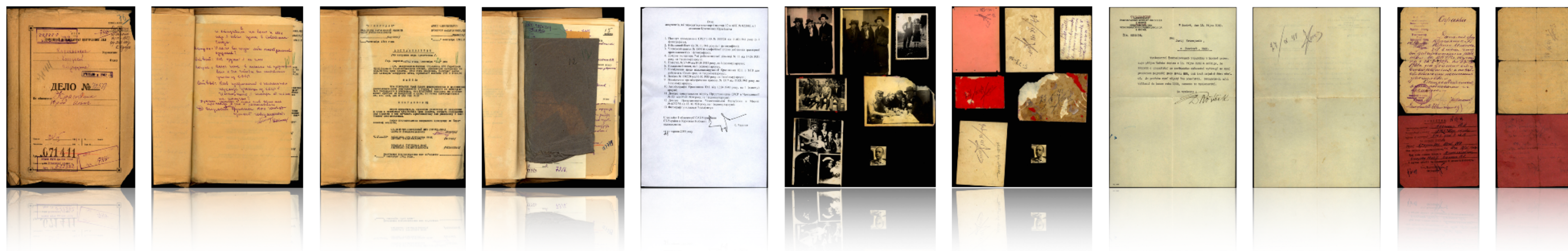
## in Relation to Czechoslovak History

<https://archivkgb.zcu.cz>



*"...thousands of Czechoslovak citizens and ethnic Czechs were victims of political repression on Soviet territory during the existence of the Soviet Union..."*

- ~1.3M scanned documents (~3.3TB of data) from Ukrainian archives
  - documents in four languages: **cs, en, uk, ru**
  - documents (+ metadata) are hierarchically sorted (Archive -> Fond -> Folder -> Document)
- ➔ our interface has two modes: 1/ data browsing 2/ instant searching by key-phrase



# Searching in Scanned Documents by Key-Phrase

## DEMO

<https://archivkgb.zcu.cz>



Digitální archiv dokumentů NKVD/KGB k československé historii Archivní pomůcky | Jak vyhledávat | O projektu CS

Hledejte v archivech... Hledejte v seznamu

V průběhu existence Sovětského svazu se tisíce československých občanů a českých krajanů staly obětmi politických represí na sovětském území. Ústav pro studium totalitních režimů v rámci projektu *Čechoslováci v Gulagu* usiluje o vyhledání, zpřístupnění a výzkum archivních pramenů sovětských bezpečnostních složek k této problematice. Webové stránky Digitální archiv dokumentů NKVD/KGB k československé historii zpřístupňují již získané a zpracované materiály - zejména osobní vyšetřovací spisy i další dokumenty z ukrajinských archivů, aby sloužily široké veřejnosti k historickému výzkumu a připomínce nevinných obětí sovětského režimu. [Více o projektu](#)

**Seznam archivů**

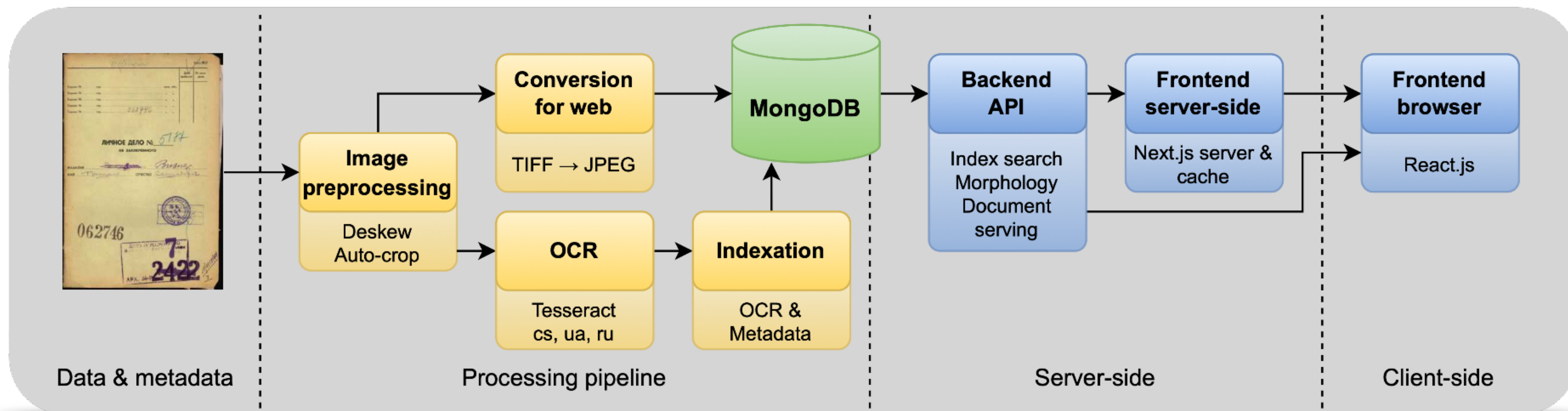
- Archiv dočasného uložení Správy Bezpečnostní služby Ukrajiny ve Lvovské oblasti (ATZ USBU Lvov) +
- Archiv dočasného uložení Správy Bezpečnostní služby Ukrajiny v Čerkasské oblasti (ATZ USBU Čerkasy) +
- Archiv dočasného uložení Správy Bezpečnostní služby Ukrajiny v Černigovské oblasti (ATZ USBU Černigov) +

**Výzkumná témata**

# Searching in Scanned Documents by Key-Phrase

## Processing pipeline

<https://archivkgb.zcu.cz>

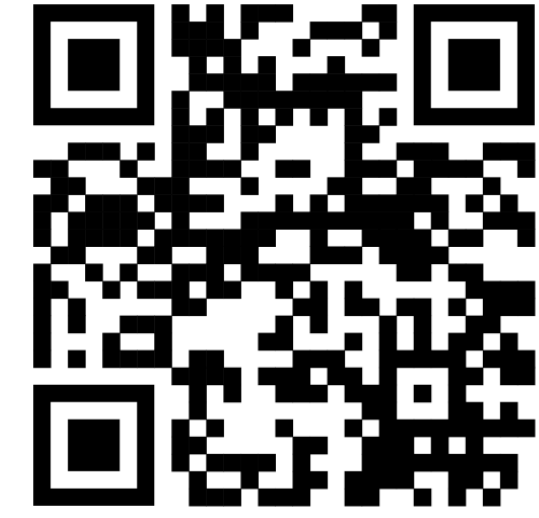


- dynamic data loading
- client-side vs. server-side rendering
- data caching

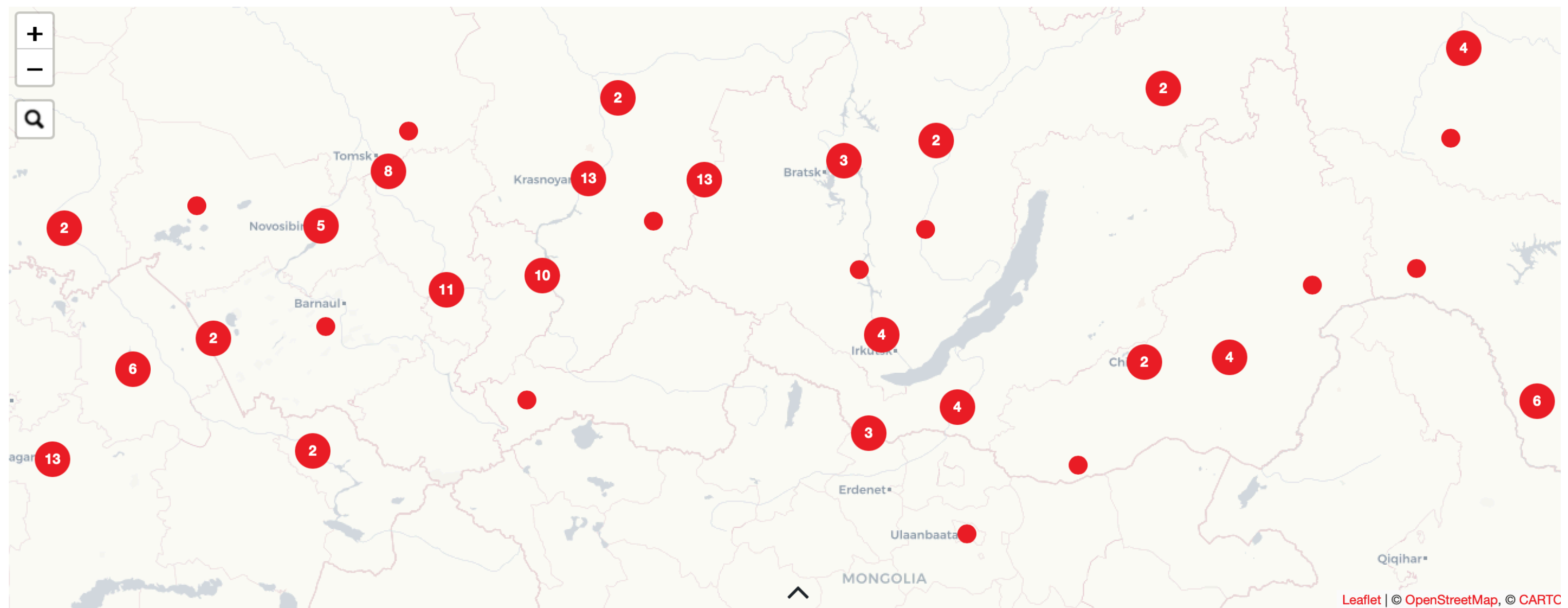
# Searching in Scanned Documents by Key-Phrase

## Next steps in the ongoing NAKI project

<https://archivkgb.zcu.cz>



- Named Entity Recognition
  - names, dates, times, places, ...
- Interactive map

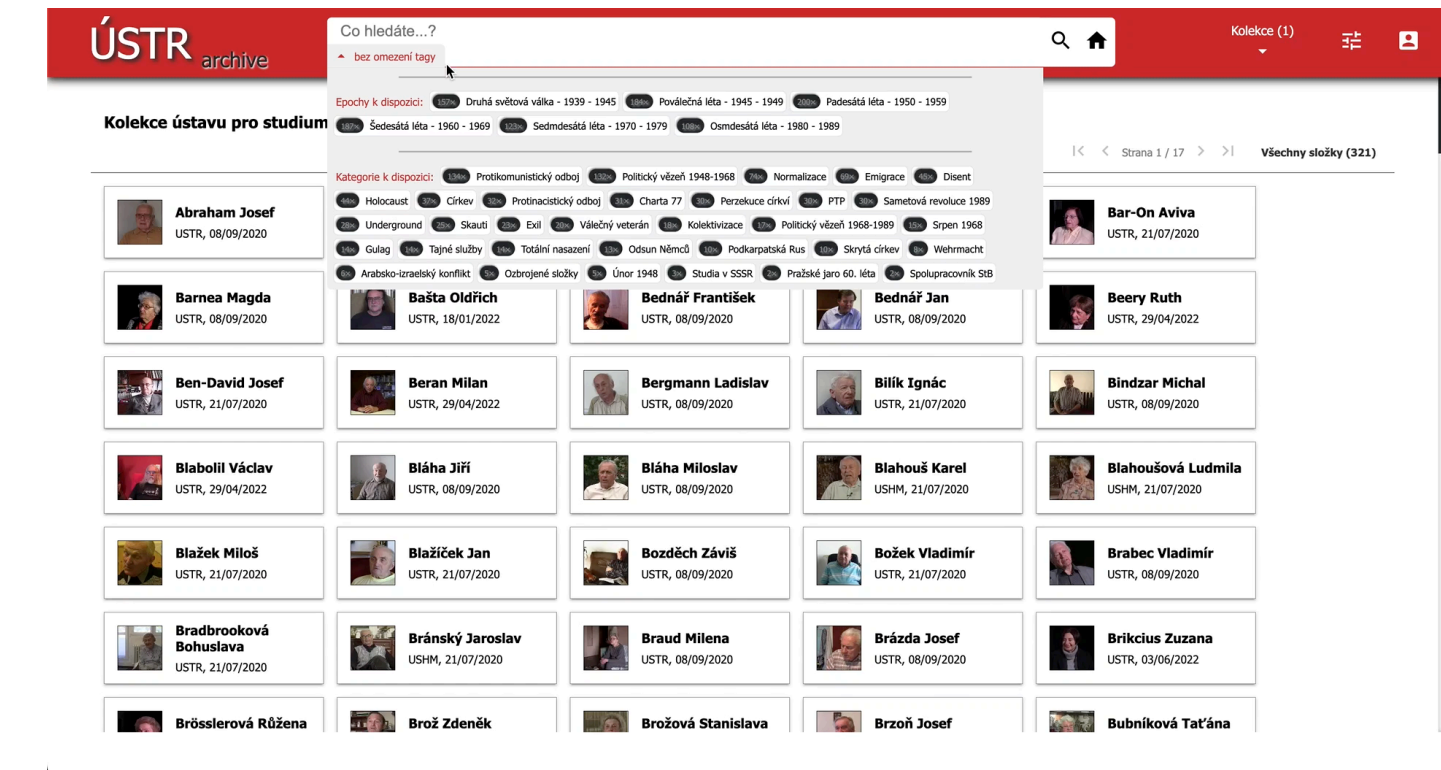


0-9 **A** B Č C Ď D E F G H I J K L M Ń N O P R S Š T ť U V W Y Z  
 Ž \*

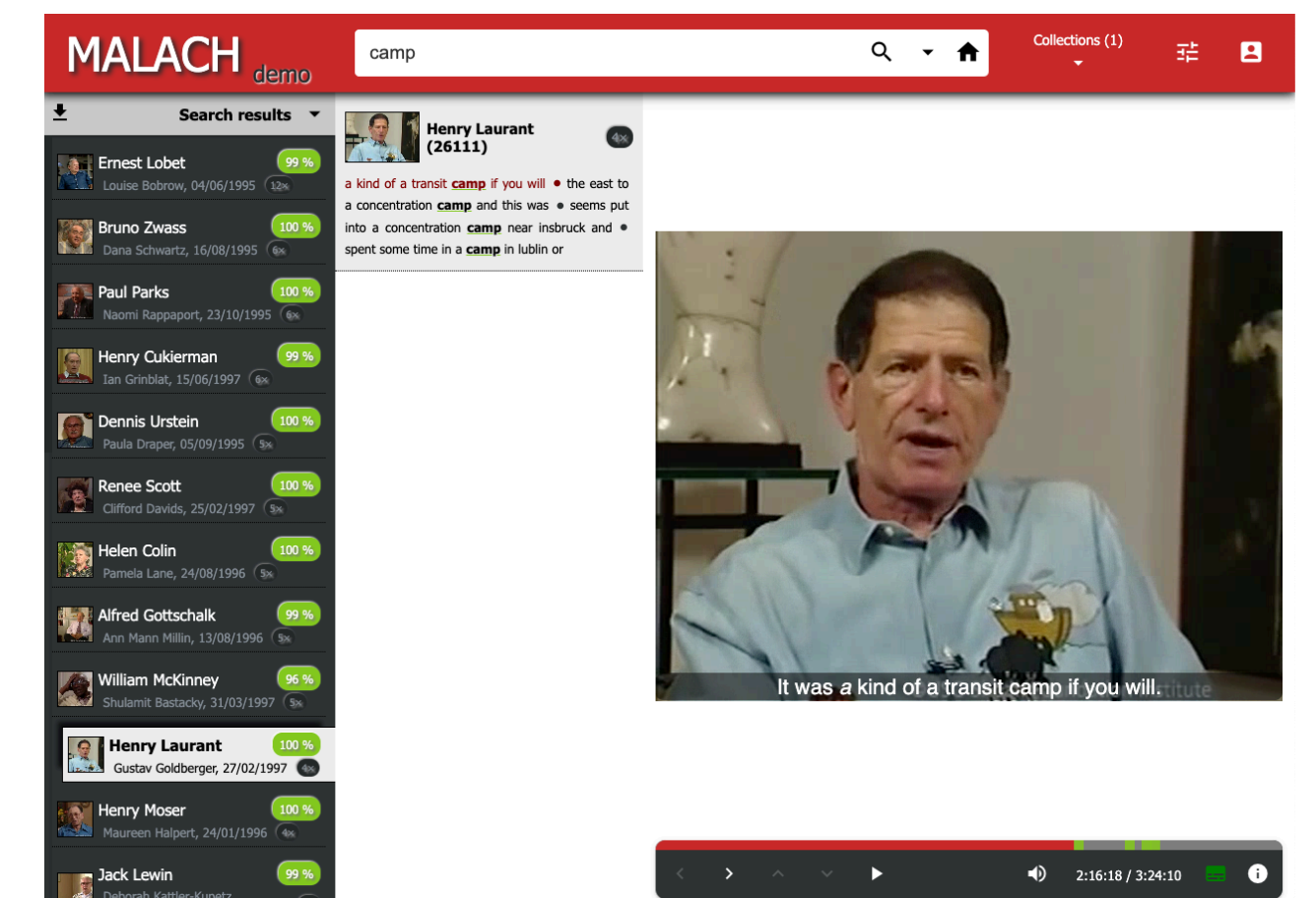
# Full-text Search in Spoken Dialogues

## Audio/video testimonies of brutal history

- The Institute for the Study of Totalitarian Regimes
  - testimonies of people persecuted by totalitarian regimes
  - 321 interviews (each approx 2,5 hours of speech)
  - interviewees have folders including relevant (scanned) documents
  - searching in recordings and scanned documents simultaneously
- Publicly available MALACH data (USC Shoah Foundation)
  - testimonies of Holocaust survivors
  - 159 interviews (each approx 3 hours)
  - collaboration with institutions on private data as well



<https://naki-ustr.zcu.cz>

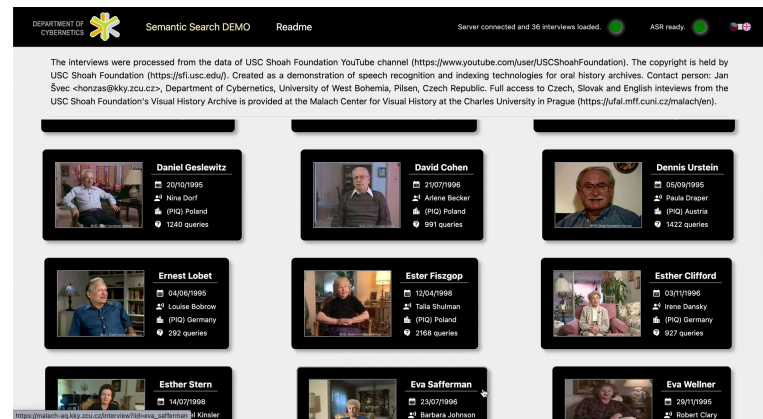


<https://malach.kky.zcu.cz>

# An Innovative Way to Interact with Large Archives

## Semantic Search

<https://malach-aq.kky.zcu.cz>



CZECH ASR ENGINE

CS => EN ONLINE TRANSLATOR

Where did you meet your wife?

Kde jste potkal vaši ženu?



**Abraham Bomba**

Description: Birth: 1913-06-06, Beuthen (Prussia, Germany); Interview: Louise Bobrow, 1996-06-14, Monticello, U.S.A.; Videographer: Daniel Liss

Date: 14/08/1996 | Gender: M | Interviewer: Louise Bobrow | Videographer: Daniel Liss

Country: Beuthen (Prussia, Germany) | City: (PIQ) Germany | Language: English | Tapes: 1

# Queries: 1576 | Length: 3:21:41 | Interview code: 18061

Video transcript: Yes, I met her in the ghetto, you met small ghetto.

Search results:

- 2:41:05 Where did the speaker meet the woman? (38%)
- 2:40:49 Where did the speaker meet the girl? (29%)
- 0:31:16 Who got married? (21%)
- 3:21:04 Where is the daughter currently living? (21%)
- 3:14:27 Where was the picture taken? (21%)
- 0:35:48 Did the speaker meet the woman? (21%)
- 2:15:27 What did the woman do outside the house? (20%)
- 2:13:10 Where does the person live? (19%)
- 2:20:58 Where did the interviewee arrive? (19%)

Semantic Similarity Model finds the most relevant indexed question

I met her in the ghetto...

CZECH TTS ENGINE

EN => CS ONLINE TRANSLATOR

future work

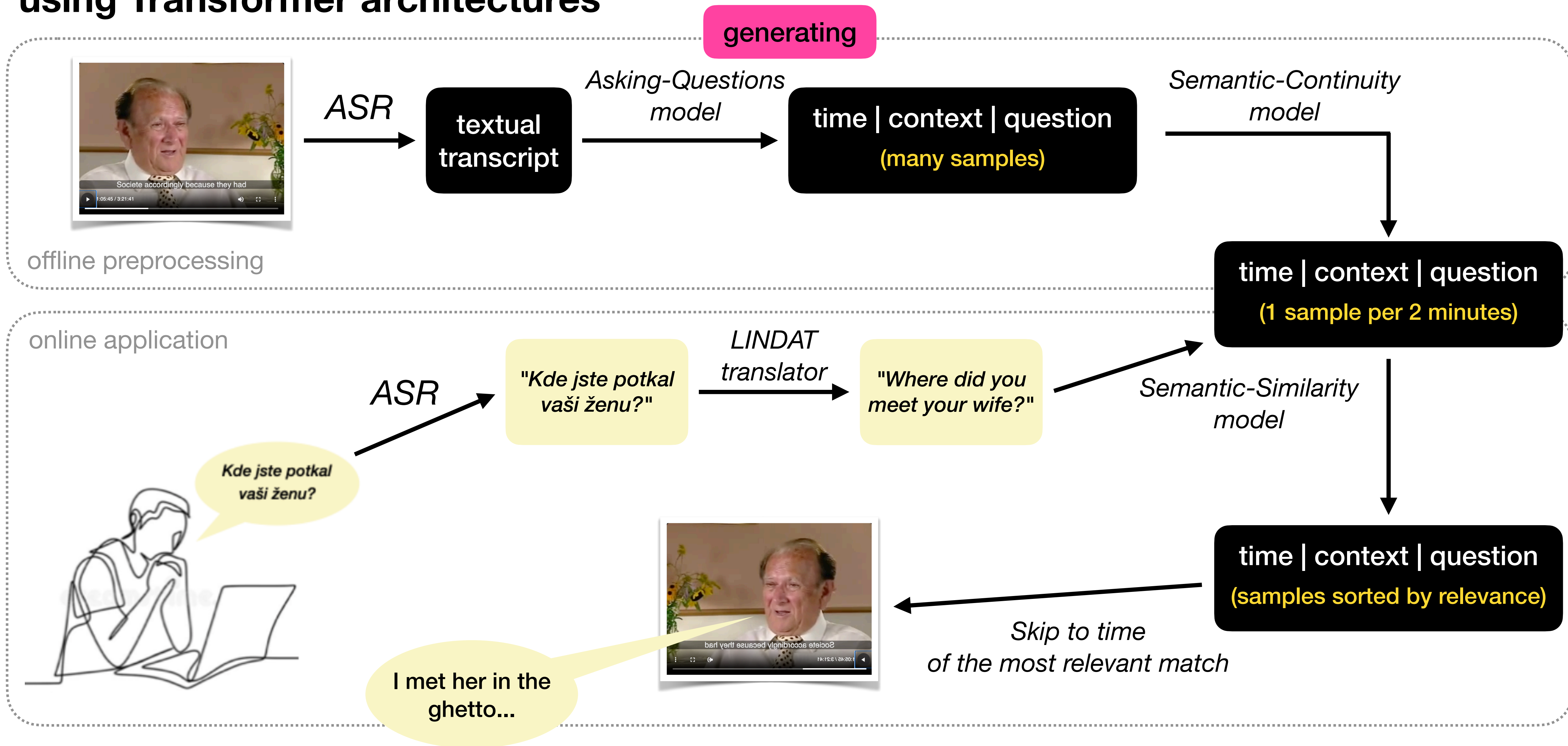
<https://malach-aq.kky.zcu.cz>





# Semantic Search in Spoken Dialogues

## using Transformer architectures



# Asking Questions Model

generating

## Generating semantically relevant questions to given contexts

- **T5** (Text-To-Text Transfer Transformer): **t5-base**

- ChatGPT API prompt

- fine-tuning on:

★ better for MALACH, as those are interviews as well

- Dataset SQUaD v2.0
- Proxy dataset generated by ChatGPT from American Life Podcasts transcripts

*You are a helpful assistant. Your task is to generate factual questions based on a provided interview context. You should aim to generate 1 to 3 general questions that can be truthfully and reasonably answered from the given context. In case the answer is not available in the context or is not mentioned in the interview, label it as <not-known>. For each generated question, please provide a straightforward answer based on the given context.*

- Context sample (American Life Podcasts)

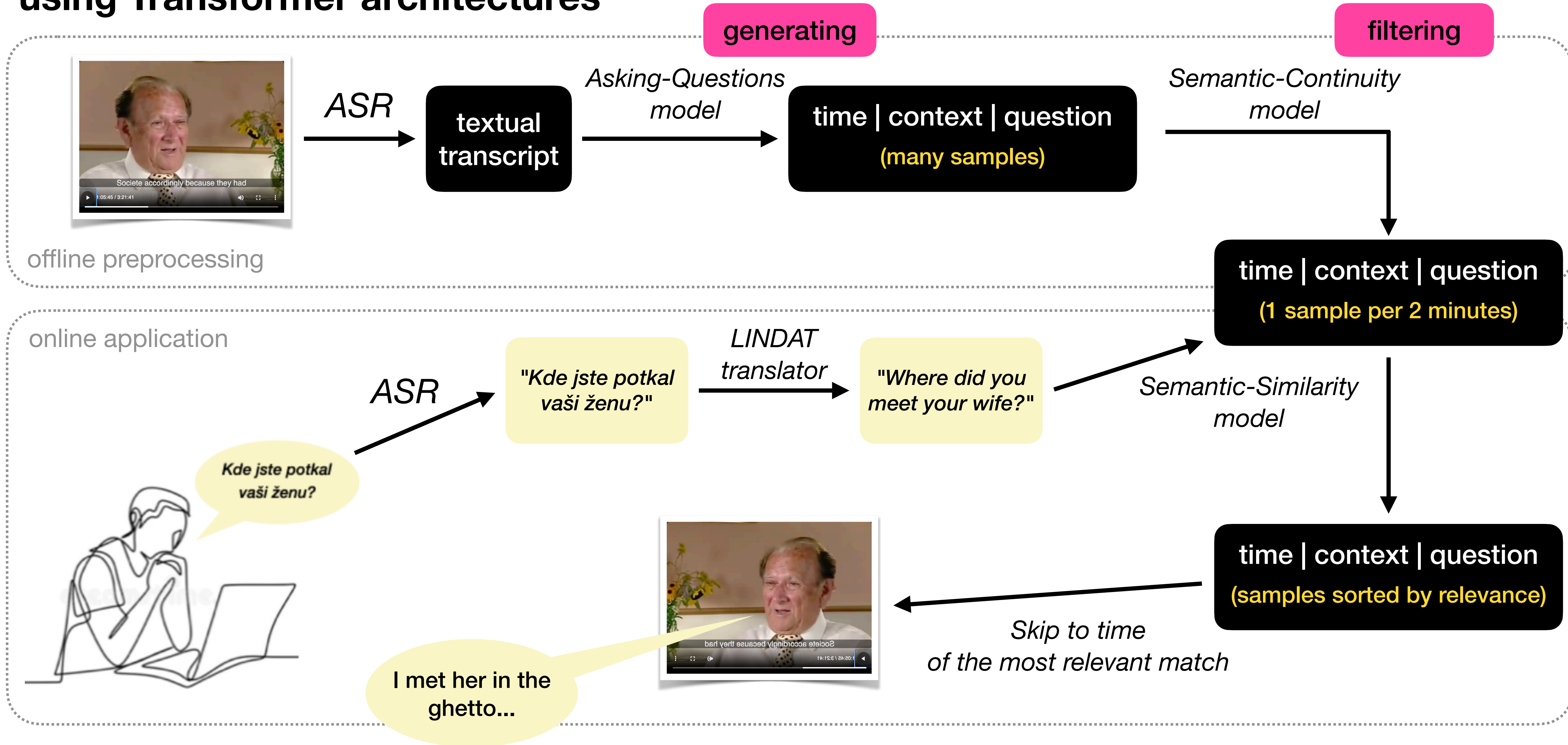
Consider, please, the role of baby talk among adults. This is speech designed specifically and entirely for cuteness, shared in private. And as I've said, our program today is about all the things that we do to actually avoid experiencing our own lives.

- Generated question-answer pairs by ChatGPT

- What is baby talk?
  - Baby talk is speech designed specifically and entirely for cuteness, shared in private.
- Is baby talk used among adults?
  - Yes, baby talk is used among adults.
- How is baby talk used among adults?
  - Baby talk is used among adults as a way to avoid experiencing their own lives.

# Semantic Search in Spoken Dialogues

## using Transformer architectures

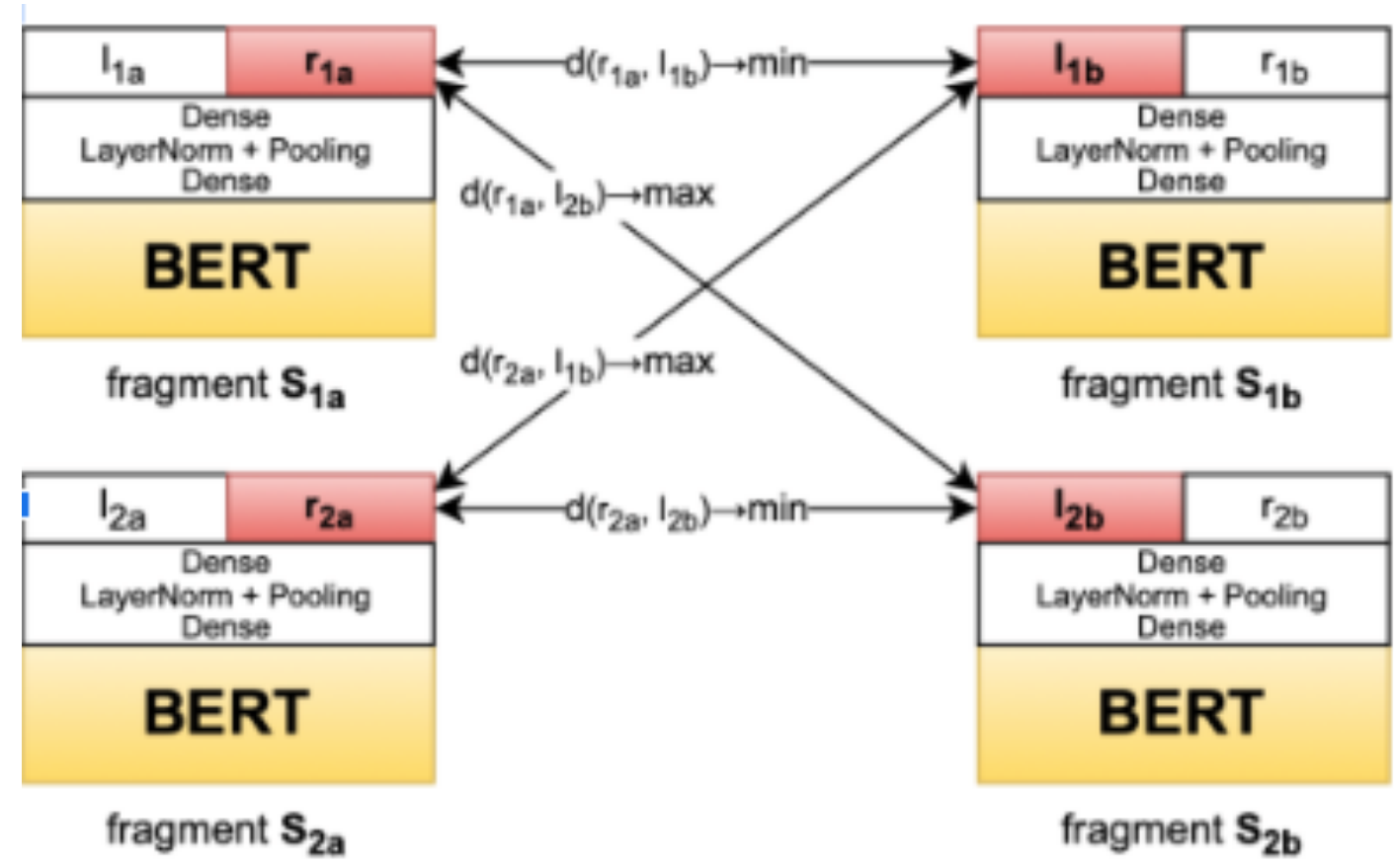


# Semantic Continuity Model

## Scoring the question relevance (semantic continuity)

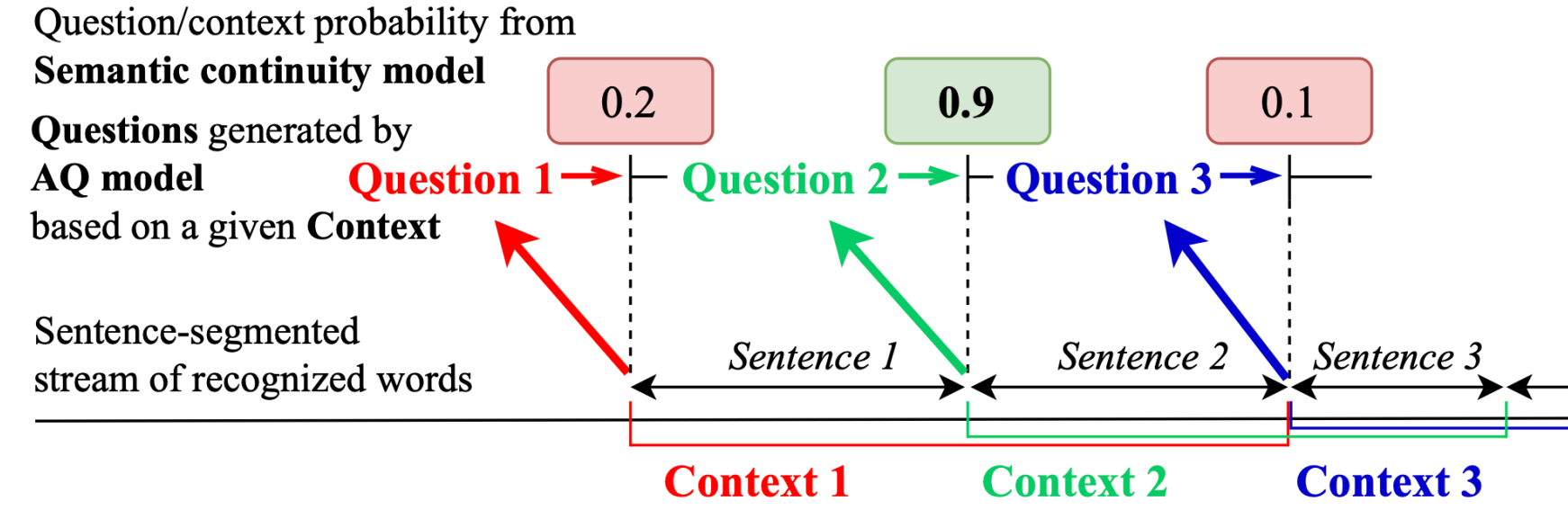
filtering

- How to score the semantic relevance??



Siamese neural network

cosine distance  
~  
the lower the better



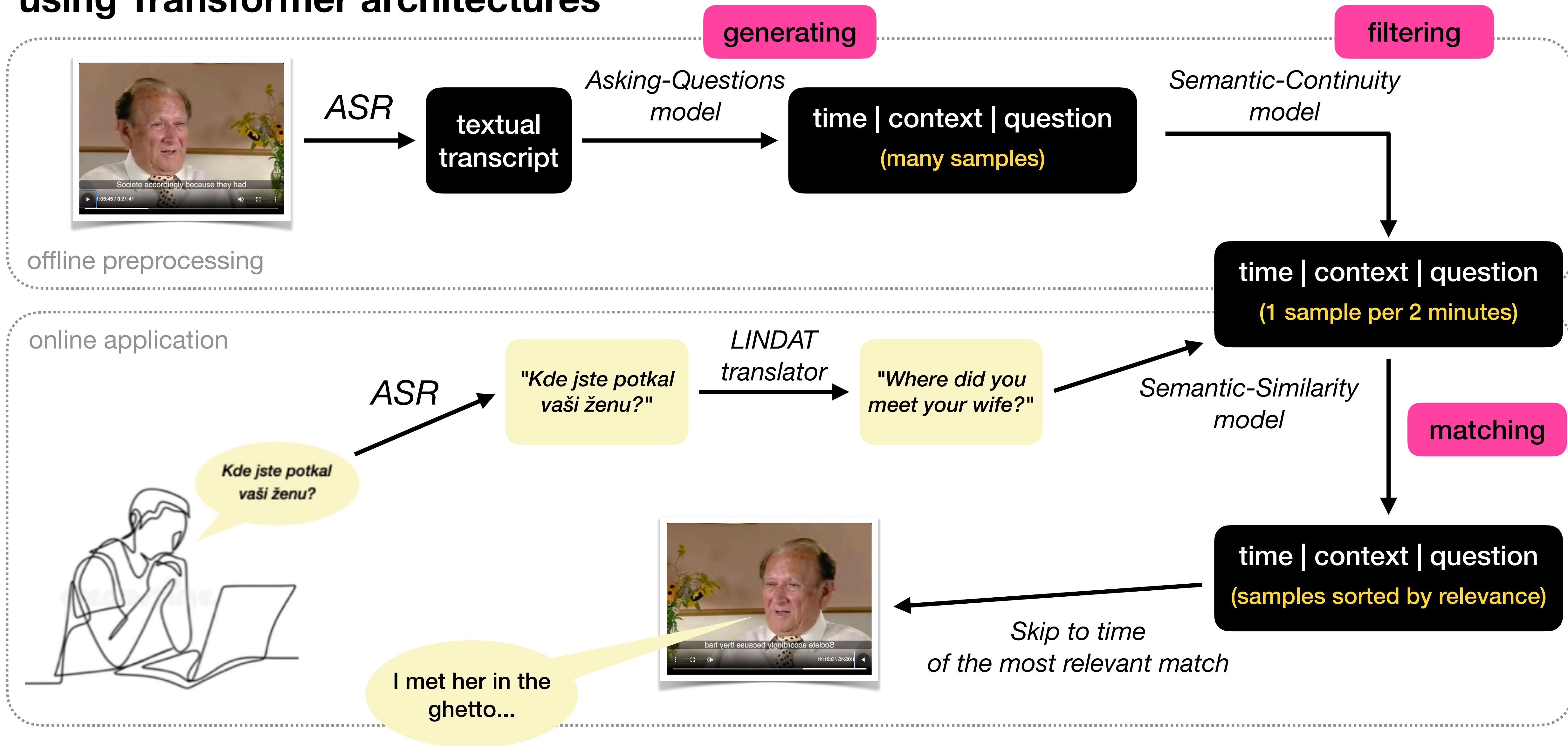
- >> Question: **What did you have for lunch?**
- >> [0.0305] For lunch we had some fish and chips.
- >> [0.0432] We did not eat at all.
- >> [0.4193] We visited the city center. Then we had fish and chips.
- >> [1.5441] We visited the city center. Then we had some lunch.
- >> [1.8320] We've been to the North London Derby. Arsenal was fantastic.

- MALACH data: 477 out of 8027 generated questions kept (5.94%)
- In average, one questions per 2 minutes of the interview

★ kinda list of TOPICS

# Semantic Search in Spoken Dialogues

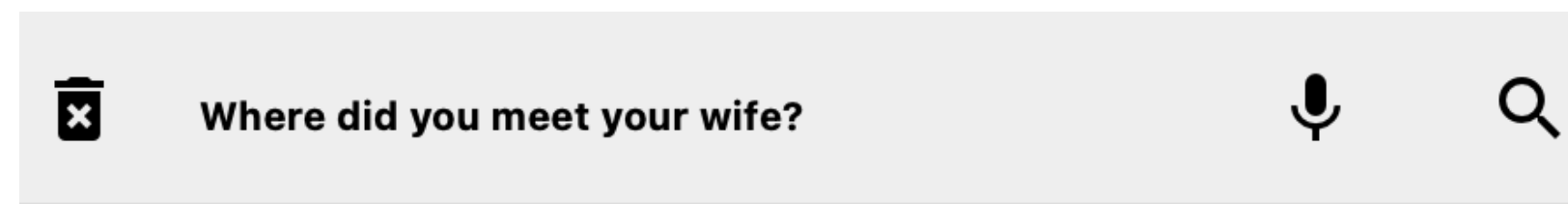
## using Transformer architectures



# Semantic Similarity Model

matching

## Scoring the semantic similarity of two sentences (questions)



- Out of the pre-generated questions, which one fits the best the user's query?
- Is there actually one? i.e. *"Is there a relevant passage in the interview?"*
- `[https://www.sbert.net/] $ pip install sentence-transformers`

★ used model:  
all-distilroberta-v1

➡ made embeddings of dim 768 and then applied the cosine distance of vectors

- Example:

>> Prompt: **How many siblings do you have?**

>> [0.7127] Do you have any brothers or sisters?

>> [1.0081] What can you tell about your family?

>> [1.1591] What was your mother's name?

>> [1.2738] What did you have for lunch?

cosine distance  
~  
the lower the better

# Next Steps

- Semantic Search Framework
  - **adding TTS** to complete the cycle of accessibility and usability
  - application on new data
  - search in entire archive (?)
- The NKVD/KGB Archive: Named Entity Recognition + interactive map
- ASR Pipeline
  - including speaker diarization
  - including other (currently) post-processing steps
- OCR: any improvements / new ideas welcome



★ general approach  
=> easily applicable for  
new data / institutions

# Take-Aways

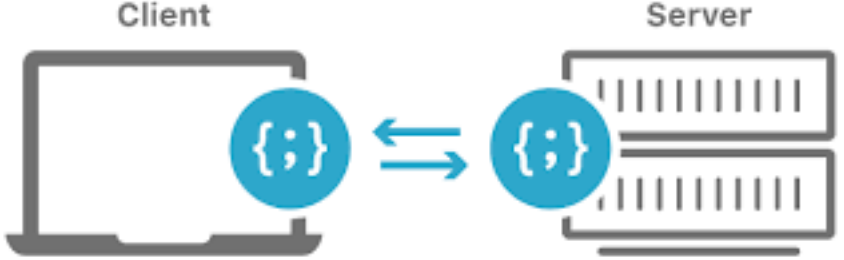
- We do not have AI, but we have #AI tools (great methods / ideas / principles / math ...)
- **Our mission** should be to find the **right applications** for these tools
- one possible application: browsing, interpreting, searching in **extensive data** collections
- regarding that, at *KKY*, we have / we make:



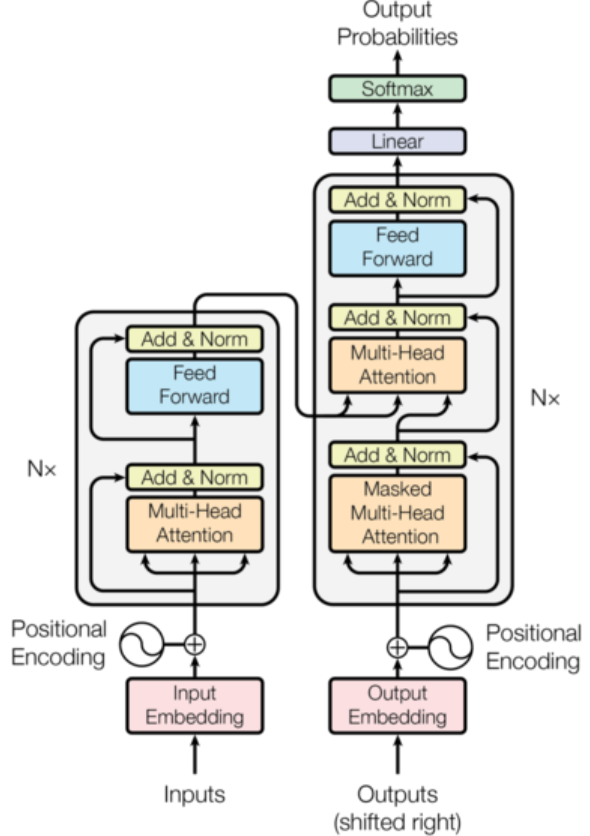
**State-of-the-Art  
ASR (en, cs, sk, de)**



**Database Indexing**



**Full-Stack Applications**



**Text Processing  
based on Latest Methods**



**Web-based  
Graphical Interfaces**



**Applications  
Tailored for Customer**



**High-Quality TTS**



# Thank you for your attention

**Q?**

**27. 10. 2023**

**Martin Bulín**  
bulinm@kky.zcu.cz

**Jan Švec**  
honzas@kky.zcu.cz

**Pavel Ircing**  
ircing@kky.zcu.cz