Automatizovaný předvýběr archiválií v kontextu českého prostředí

Pavlína Nimrichtrová

27. listopadu 2025

Národní archiv, Univerzita Karlova, katedra PVH a archivnictví

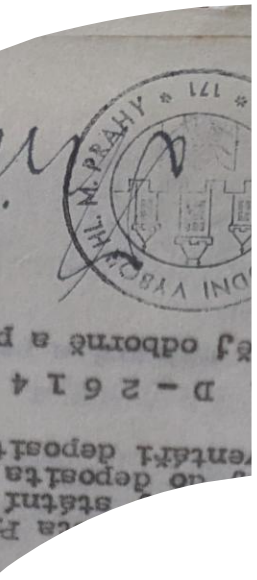
Proč?

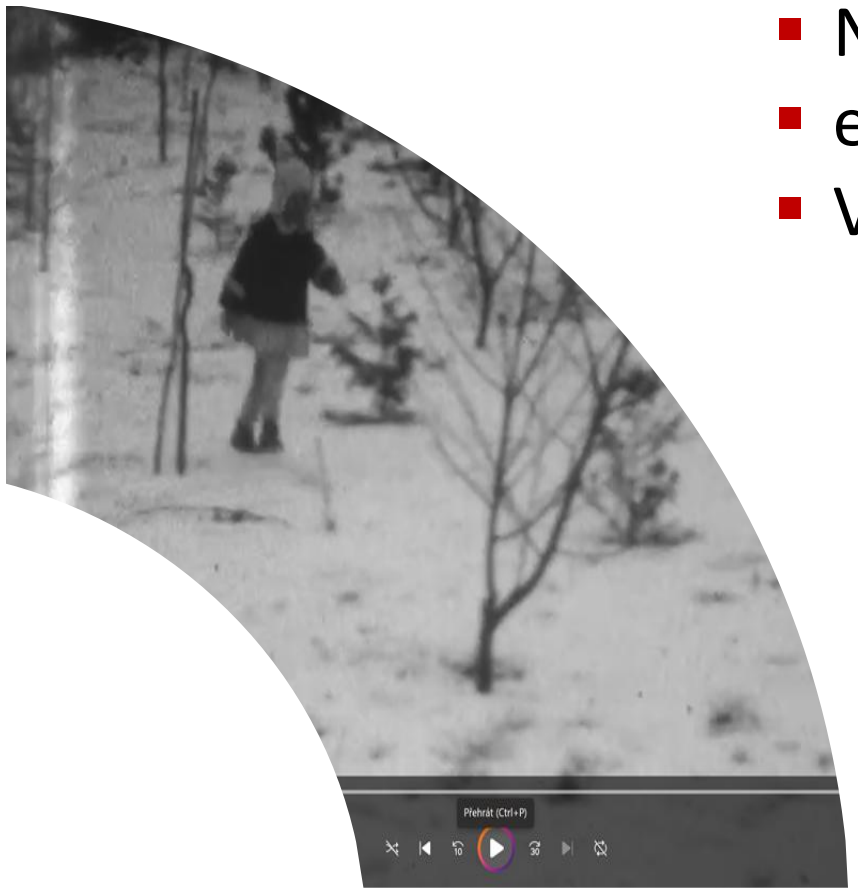
- Projekt předvýběru (2024-)
- Národní archiv - garantem výběru archiválií – Národní archivní portál (eSkartace, Výběr z volných souborů, výběr z IS spadajících pod § 3a)
- Opravdu je to potřeba – množství, přesnost, „tekutost“

Co?

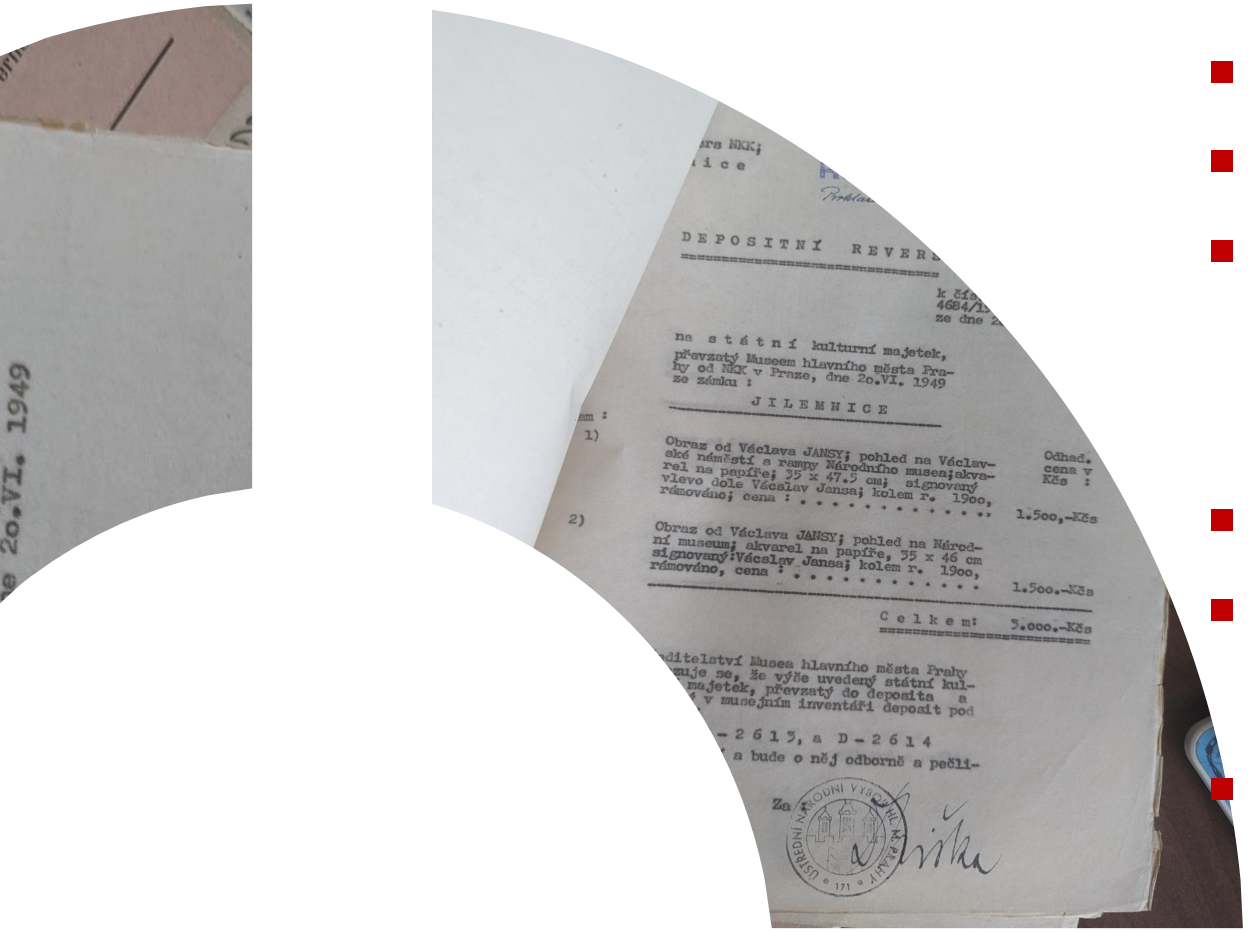
- Naivní představy
- (Dnešní postup – 90% manuální (archivář vybírá, spisová služba – přidělení parametrů – dosud ručně))
- Rozhodnutí se přiřadí samo – archivář jenom schválí (-human in the loop-)
- Inspirace – Národní archiv UK, Národní archiv Singapur, …

Kde to potřebujeme?





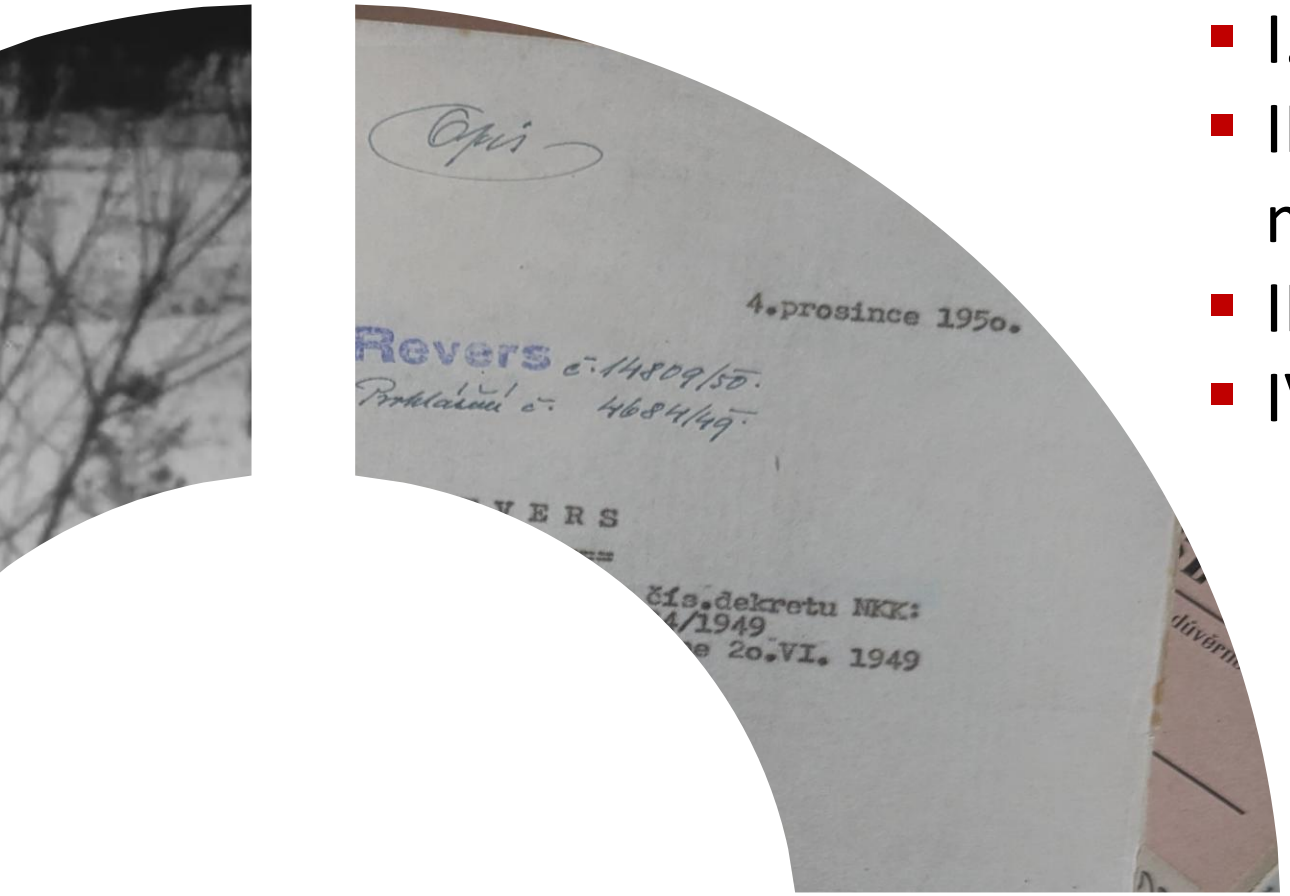
- Národní archivní portál
- eSSL
- Výběr u soukromých osob, původců



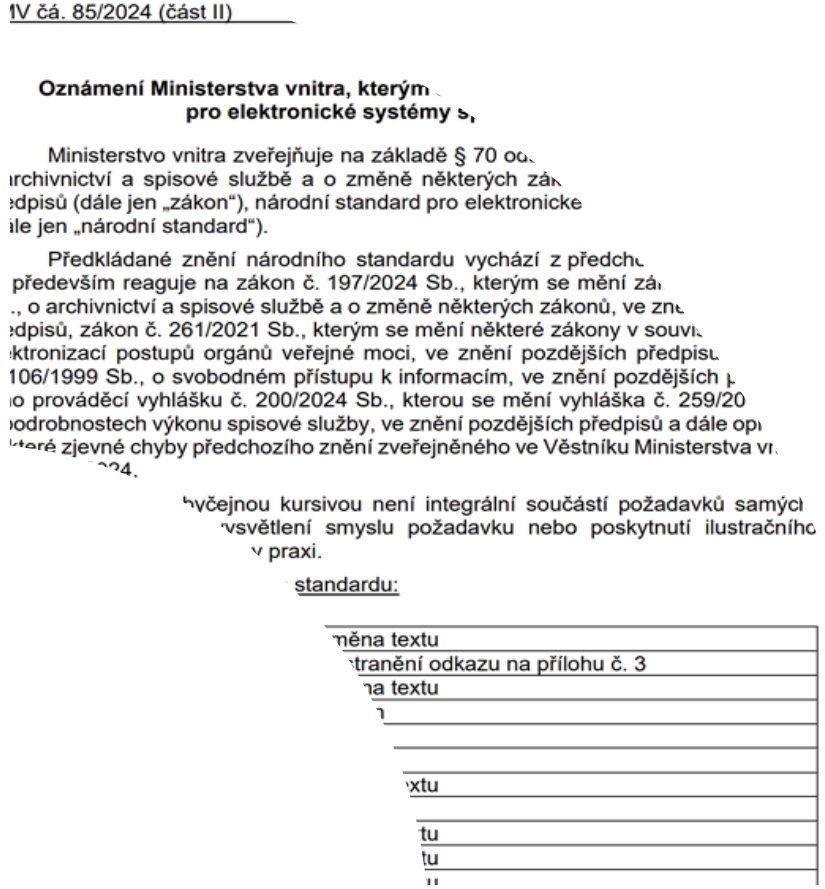
Co to znamená?

- Nutnost stanovit use case
- Jak budou vypadat vstupy a výstupy
- Jaká můžeme použít trénovací, testovací data (co je k dispozici, na co máme nárok)
- Jaké přístupy půjde použít
- Jaké technologie, nástroje, modely půjde použít
- Jak průběžně kontrolovat a zlepšovat výsledky

Use Case



- I. Výběr podle seznamu – In progress
- II. Výběr ze SIP balíčků (tzn. existují metadata) – TO DO
- III. Výběr ze souborů – In progress
- IV. DocuMate App In progress



IV čá. 85/2024 (část II)

Oznámení Ministerstva vnitra, kterým
pro elektronické systémy s,

Ministerstvo vnitra zveřejňuje na základě § 70 od
rchivnictví a spisové službě a o změně některých zá
dpisů (dále jen „zákon"), národní standard pro elektronické
le jen „národní standard").

Předkládané znění národního standardu vychází z předch
především reaguje na zákon č. 197/2024 Sb., kterým se mění zá
., o archivnictví a spisové službě a o změně některých zákonů, ve zn
dpisů, zákon č. 261/2021 Sb., kterým se mění některé zákony v souvi
ktronizací postupů orgánů veřejné moci, ve znění pozdějších předpis
106/1999 Sb., o svobodném přístupu k informacím, ve znění pozdějších
o prováděcí vyhlášku č. 200/2024 Sb., kterou se mění vyhláška č. 259/20
odrobnostech výkonu spisové služby, ve znění pozdějších předpisů a dále op
teré zjevné chyby předchozího znění zveřejněného ve Věstníku Ministerstva v
24.

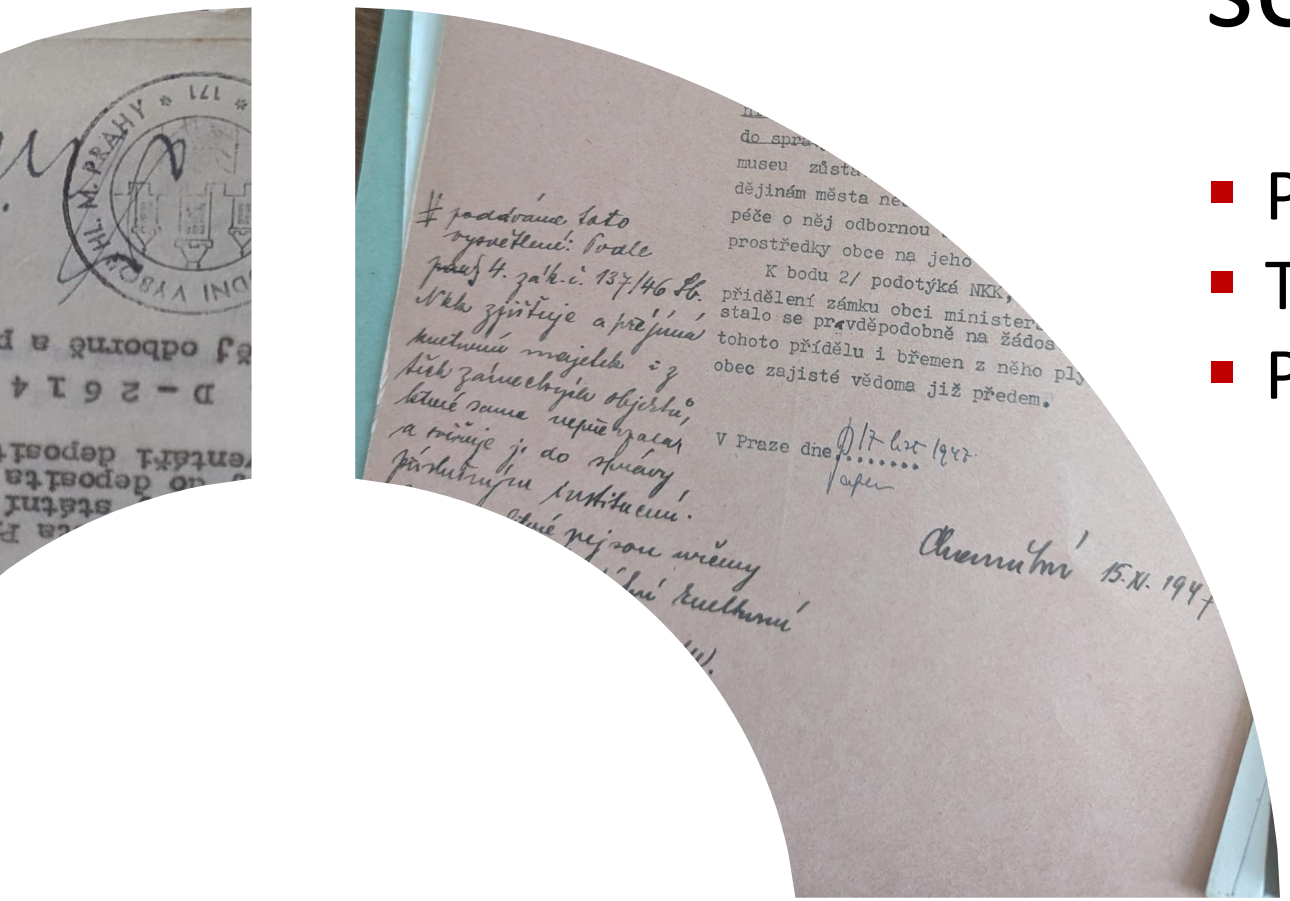
včejnou kursivou není integrální součástí požadavků samýc
vsvětlení smyslu požadavku nebo poskytnutí ilustračního
v praxi.

standardu:

měna textu
tranění odkazu na přílohu č. 3
na textu
xtu
tu
tu

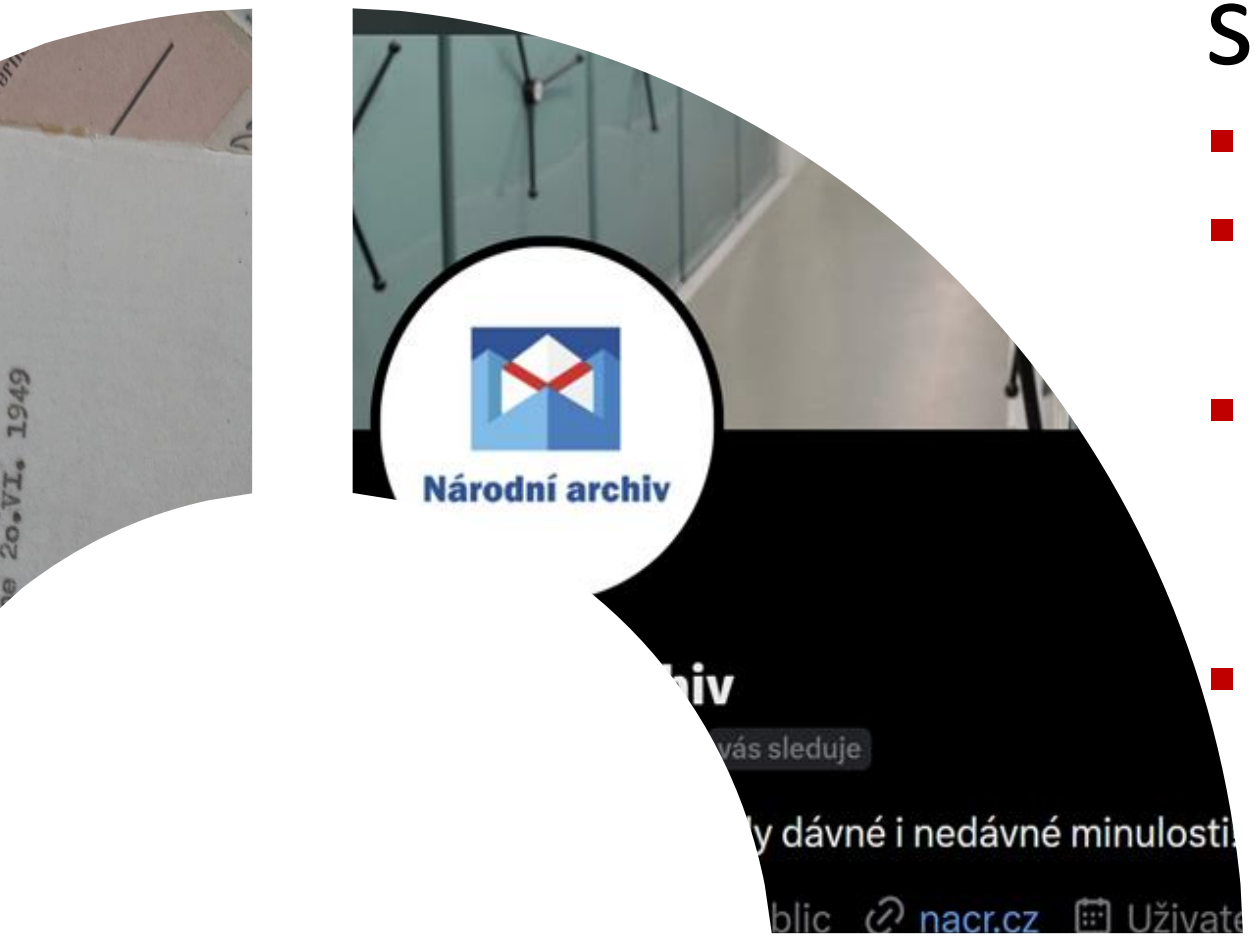
Scénář I. – výběr dle seznamu pro NArP

- Model mBERT
- Workflow - stáhnout excel z ESK portálu
  - pomocí skriptu přiřadit rozhodnutí
  - nahrát zpět do portálu



Scénář I. – výběr dle seznamu pro NArP

- Příprava dat
- Trénování modelu
- Posouzení

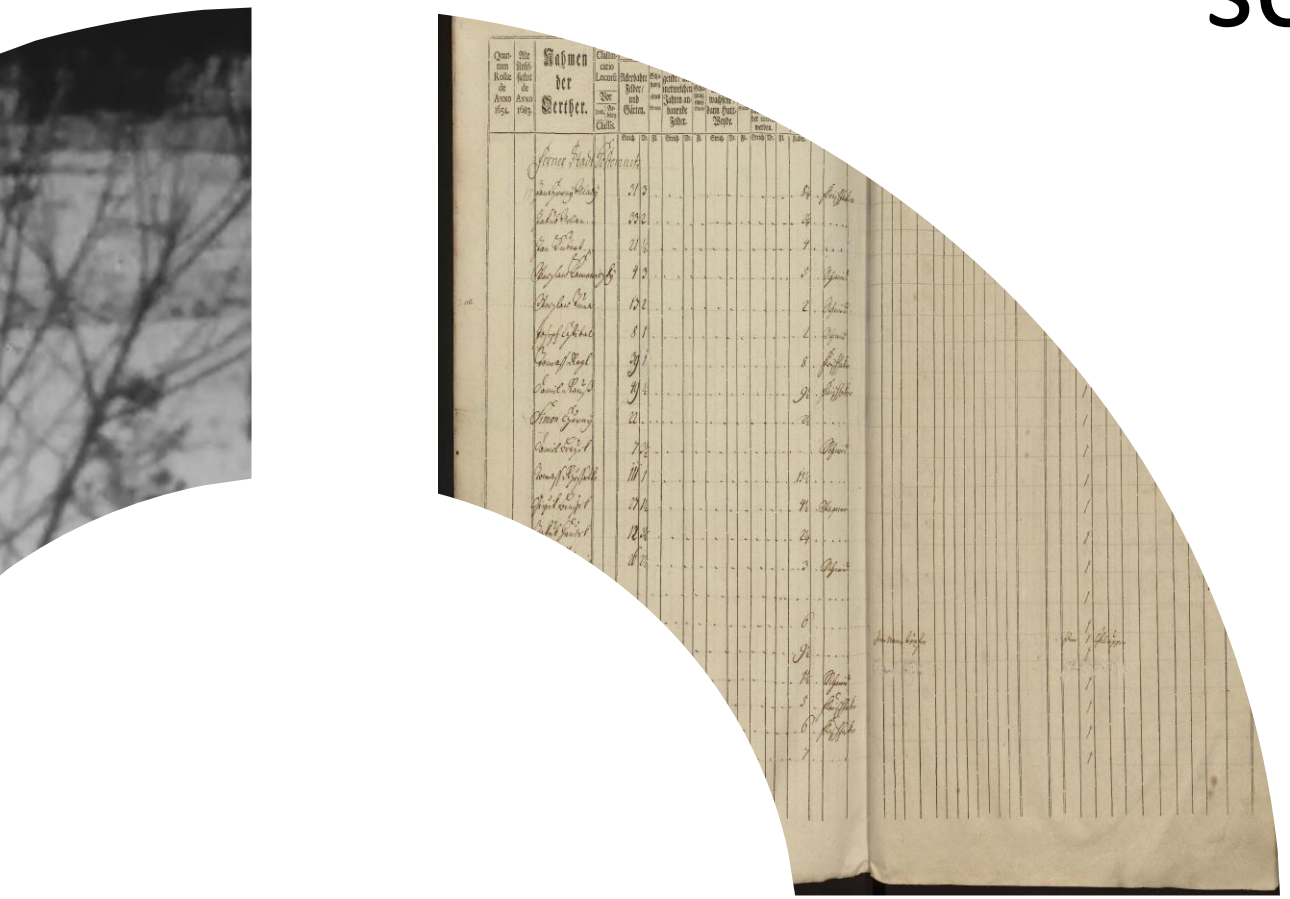


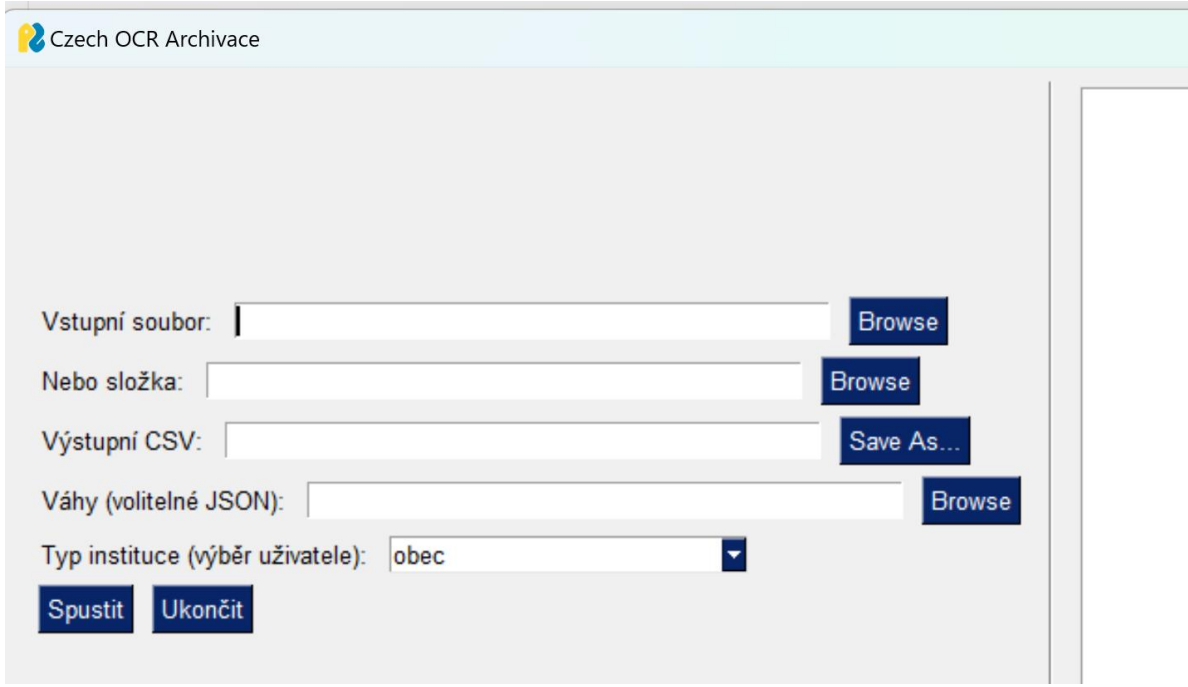
Scénář I. – výběr dle seznamu pro NArP

- Výsledky zatím nic moc
- Potřeba odladění klasifikace, hodnocení – formou vhodných tréninkových dat
- Přidat nové parametry

- Další kroky (samostatná apka, implementace do portálu…)

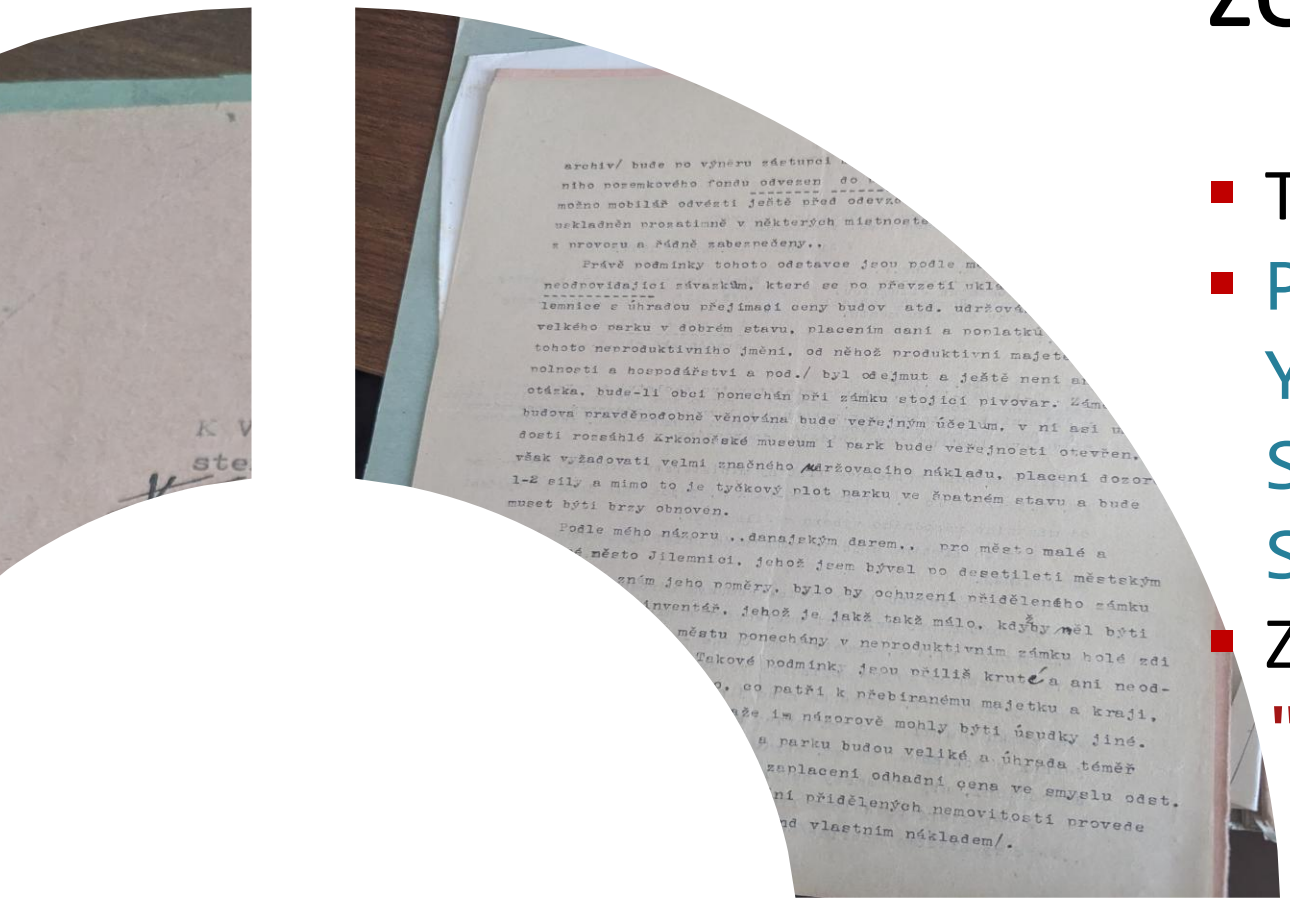
Scénář III. – výběr ze souborů





Scénář III. – výběr ze souborů

- Vstup – soubory, zadání parametrů
- Úkoly – OCR, extrakce textu
  rozpoznání jmen, míst, institucí, klíčových slov
  rozpoznání typu dokumentu
  sumarizace v 1 větě,
  vyhodnocení (body)
- Výstup – seznam souborů s hodnotou A/S/V/X



Scénář III. – výběr ze souborů

- Tesseract
- `PySimpleGUI`, `pytesseract`, `YAKE(Yet Another Keyword Extractor)`, `SUMY (Extraktivní sumarizace)`, `Stanza (NER)`
- Zatím pouze `".jpg"`, `".jpeg"`, `".png"`

Scénář III. – výběr ze souborů

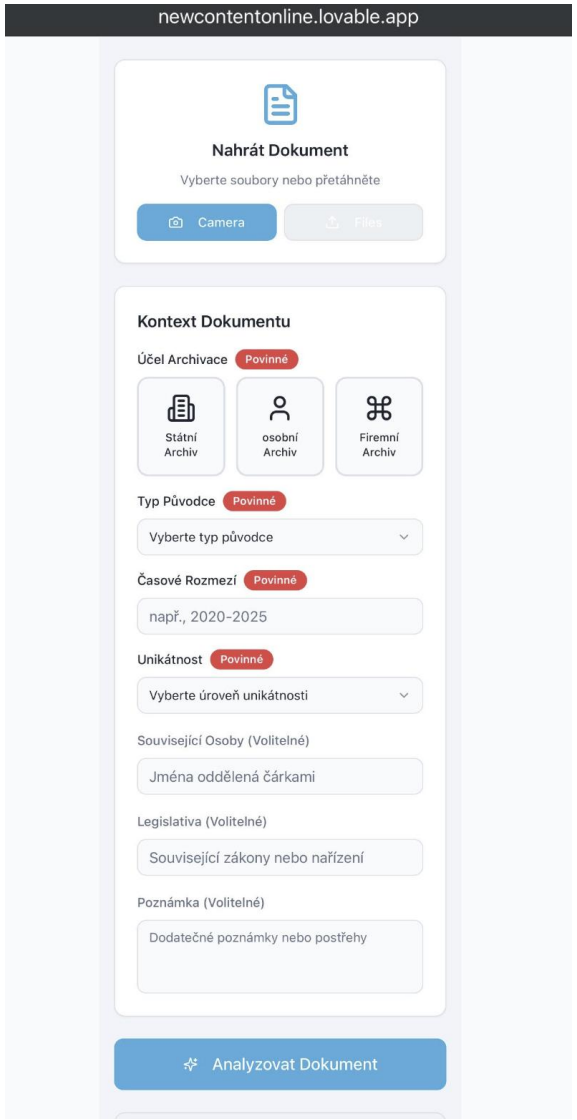
- Nastavení OCR (např. ořez razítek, narovnání)
- Jiná metoda sumarizace (RAG sumarizace)
- Klíčová slova (např. jiná stop slova, blacklist)
- Další formáty vstupu
- Jiné nastavení klasifikace (hodnoty skóre)

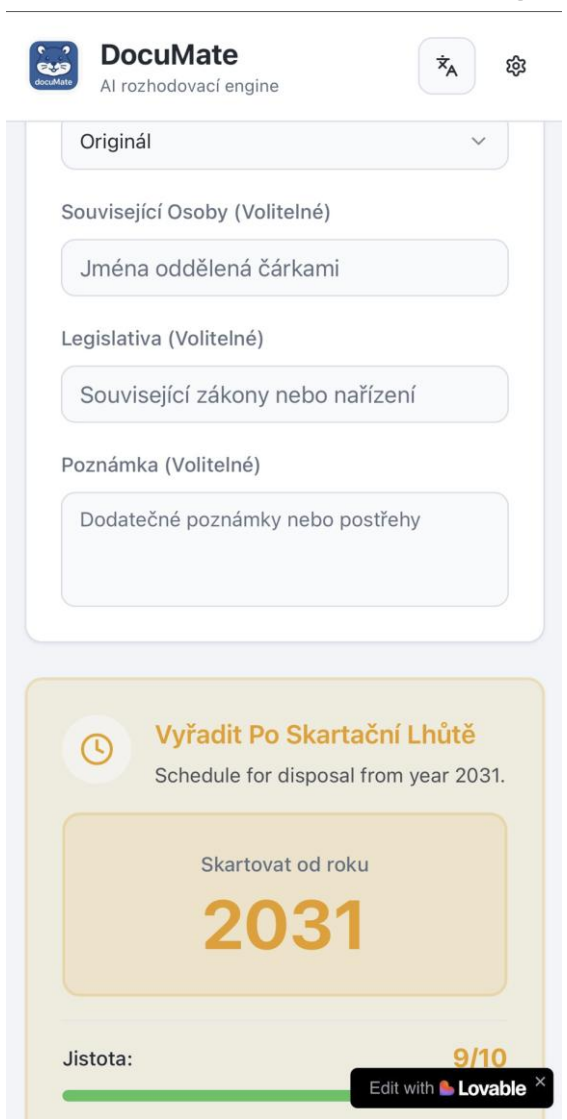
Scénář IV. – DocuMate App

- Testování kritérií a parametrů
- Testování možnosti crowdsourcingu

- Posouzení jednotlivých souborů

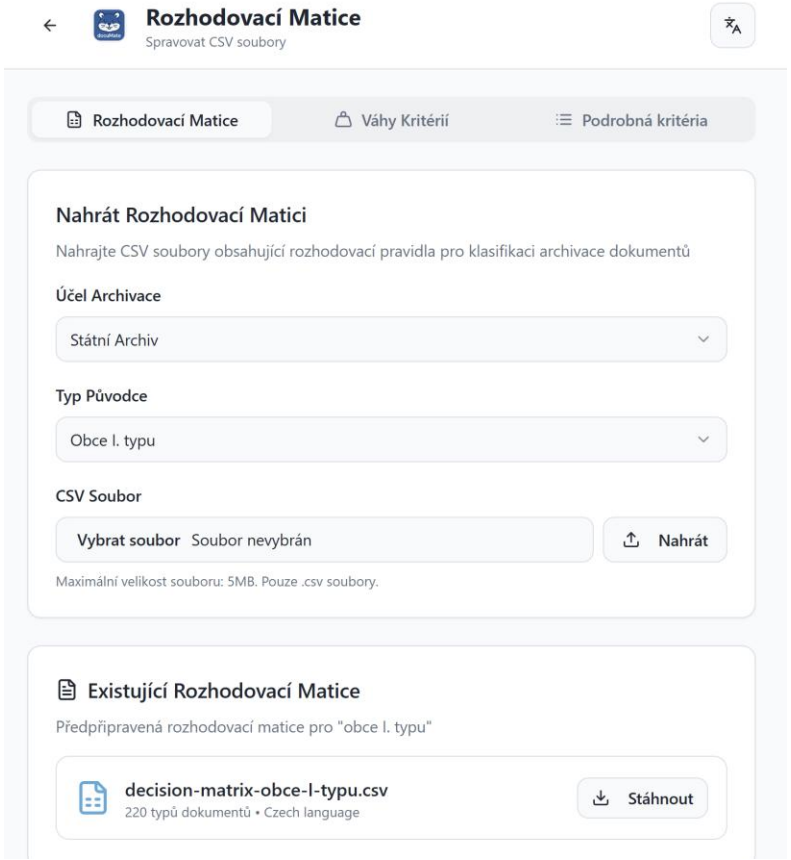
Scénář IV. – DocuMate App

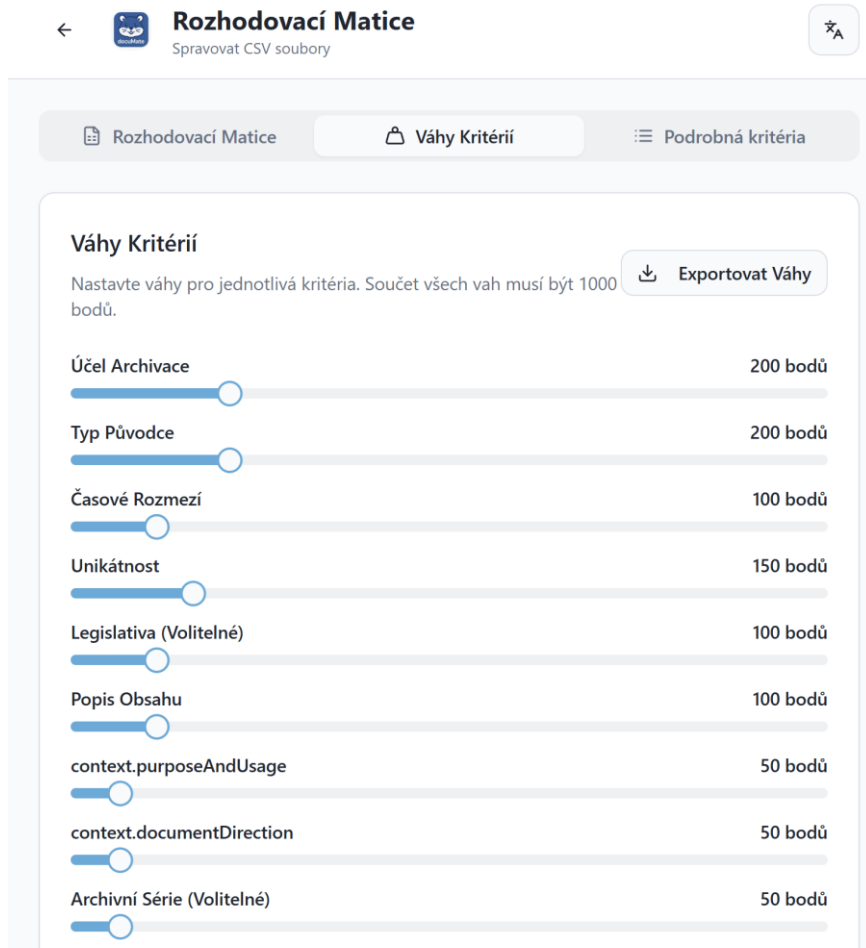


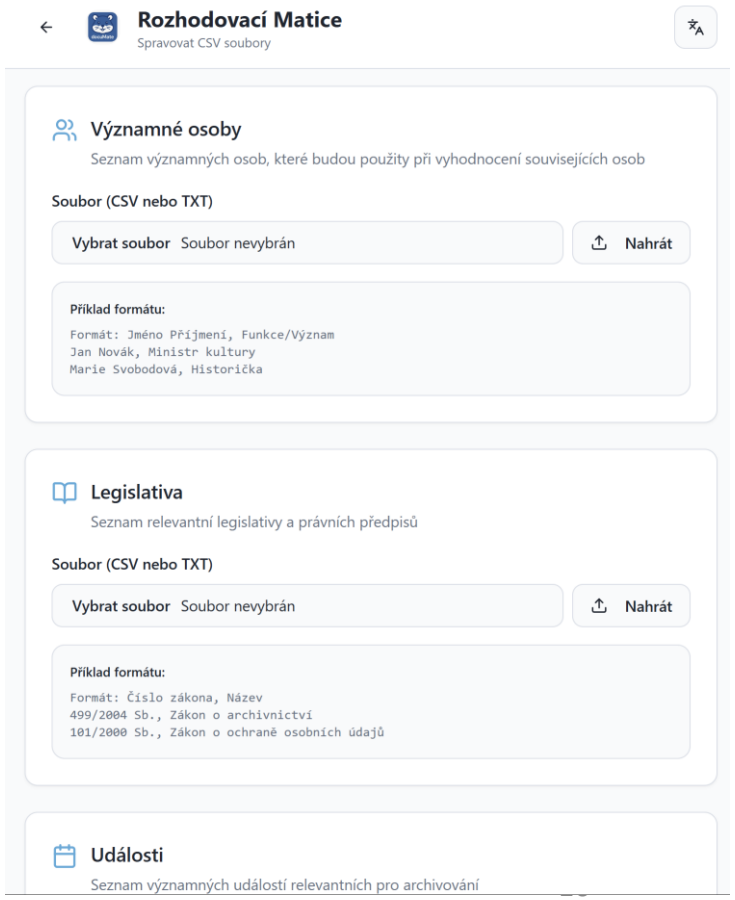


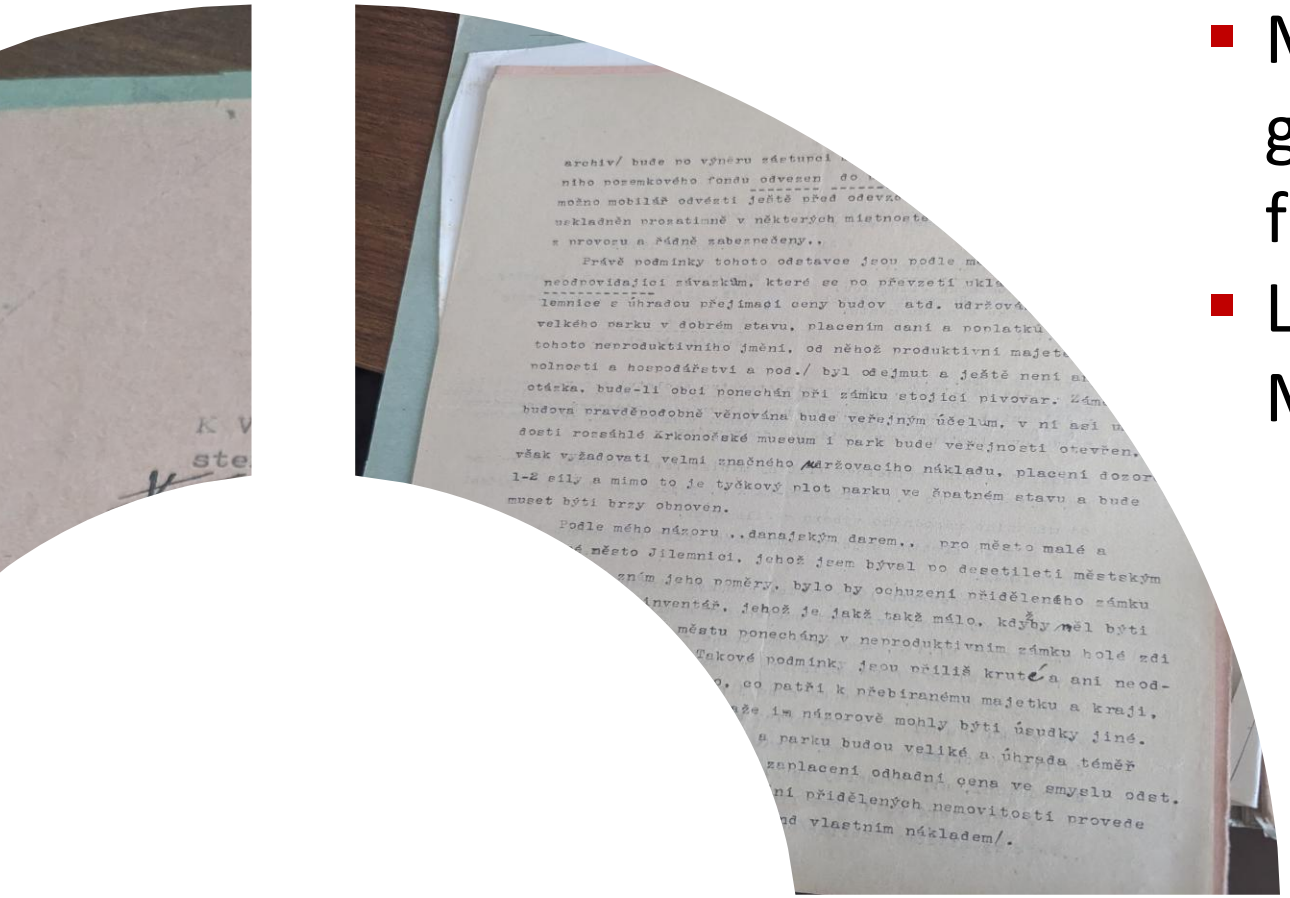
Scénář IV. – DocuMate App

- Testování kritérií a parametrů









Scénář IV. – DocuMate App

- Testování kritérií a parametrů
- Možnost využití on cloud – LLM google/gemini-2.5-flash nebo google/gemini-2.5-pro
- Lze přenést on prem (využití Llama 3.3, Mistral, Gemma)

Děkuji za pozornost

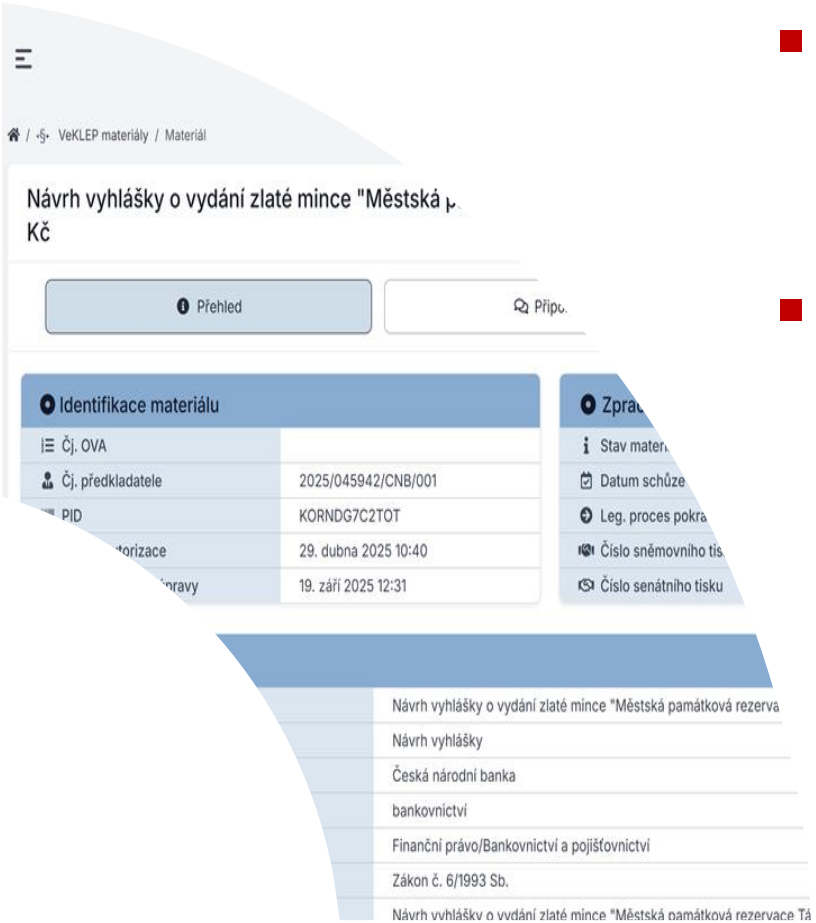
Pavlína Nimrichtrová

pavlina.nimrichtrova@na.gov.cz

27. listopadu 2025

Národní archiv, Univerzita Karlova, katedra PVH a archivnictví

Klasifikace, na základě čeho rozhodovat?



- Anketa mezi archiváři - Intuice, selský rozum, zkušenost
- Využít:

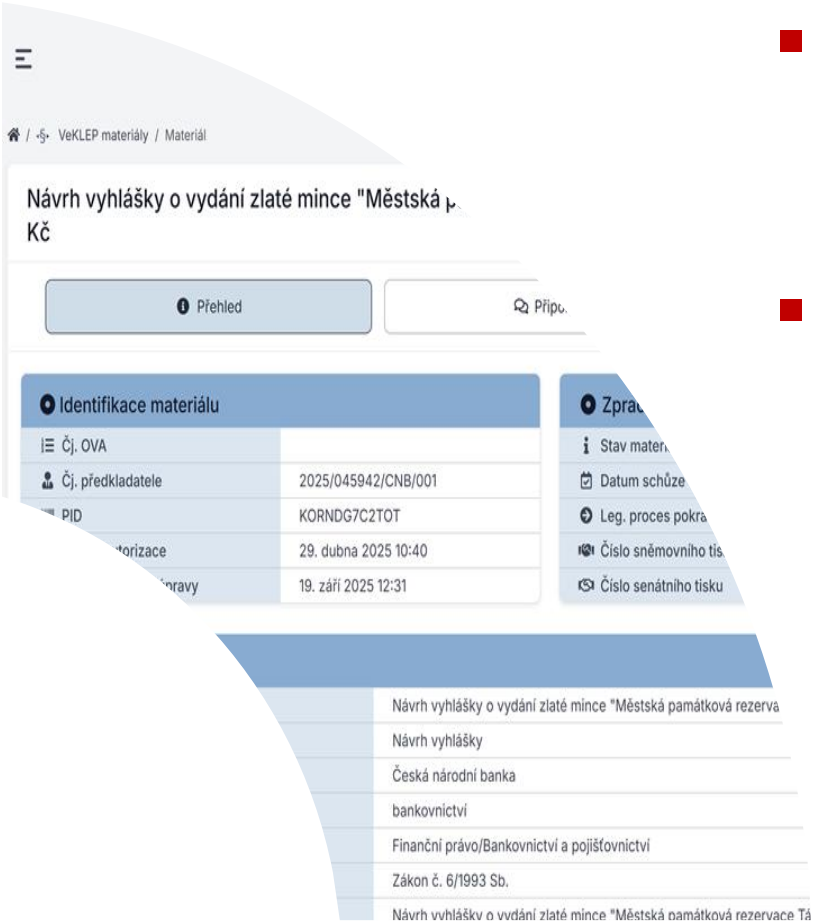
  spisové plány,
  typové skartační rejstříky,
  archivní pomůcky (EAD3),
  provedené výběry archiválií
  archivní entity – Centrální archivní modul

Klasifikace, na základě čeho rozhodovat?

- **Zákon 499/2004 je to dle §2 písm. f) archiválií** (je) takový dokument, který byl vzhledem k době vzniku, obsahu, původu, vnějším znakům a trvalé hodnotě dané politickým, hospodářským, právním, historickým, kulturním, vědeckým nebo informačním významem vybrán ve veřejném zájmu k trvalému uchování a byl vzat do evidence archiválií….

- **§5 písm (1)** Podle doby vzniku… **(2)** Podle obsahu se za archiválie příslušným archivem vybírají dokumenty, které mají trvalou hodnotu danou jejich politickým, hospodářským, právním, historickým, kulturním, vědeckým nebo informačním významem; k výběru musí být vždy předloženy dokumenty uvedené **v příloze č. 2** k tomuto zákonu. **(3)** Podle původu…**(4)** Podle vnějších znaků …

Klasifikace, na základě čeho rozhodovat?



- Anketa mezi archiváři - Intuice, selský rozum, zkušenost
- Využít:

  spisové plány,
  typové skartační rejstříky,
  archivní pomůcky (EAD3),
  provedené výběry archiválií
  archivní entity – Centrální archivní modul