

# Automatizovaný předvýběr archiválií v kontextu českého prostředí

Pavλίna 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“

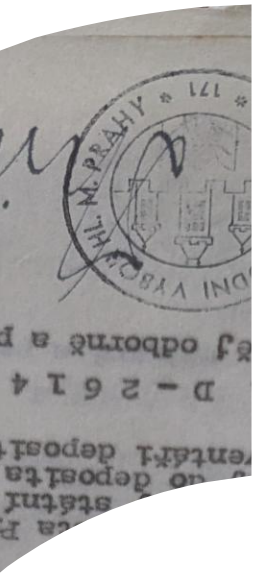
# Co?

- 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, ...



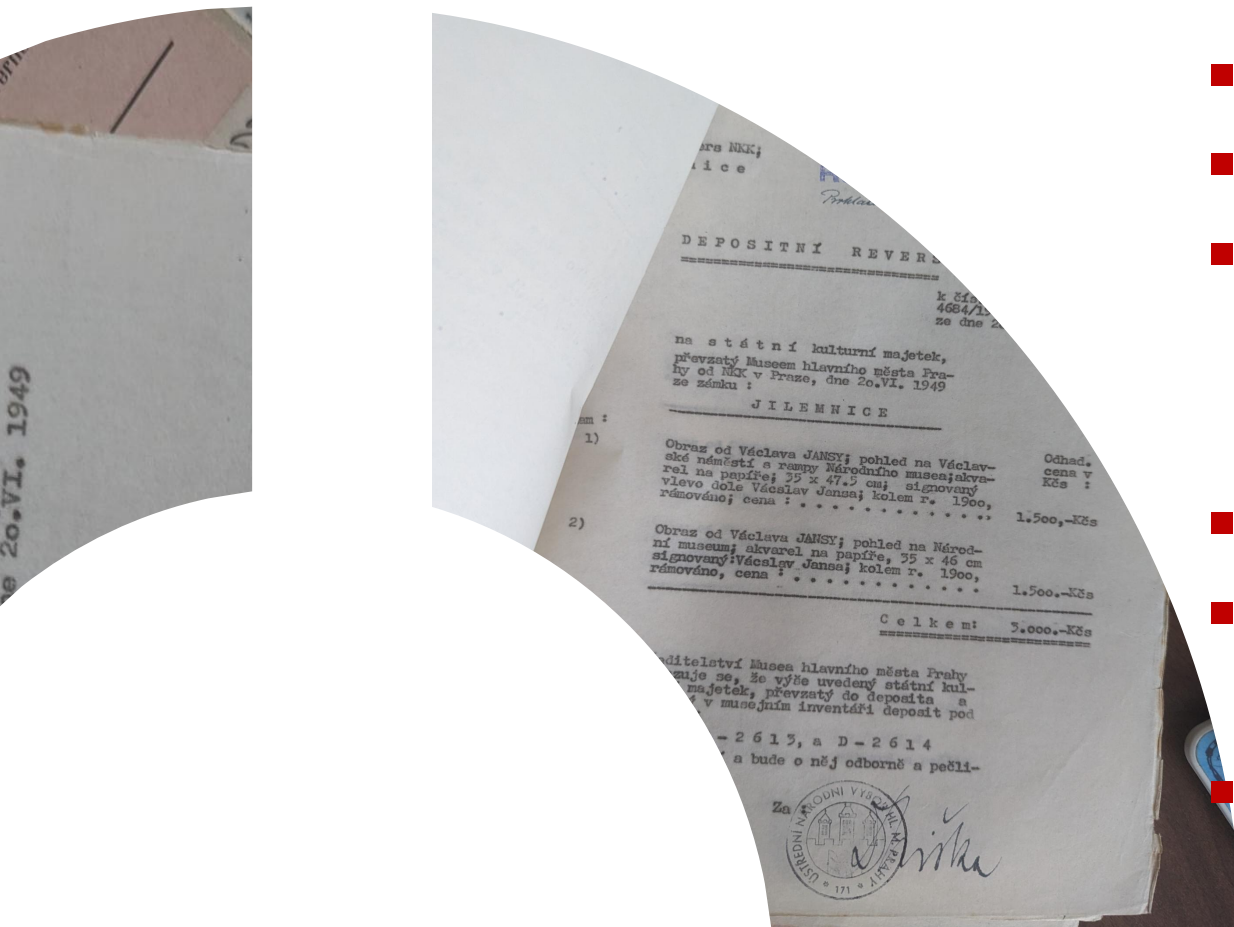
# Kde to potřebujeme?

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



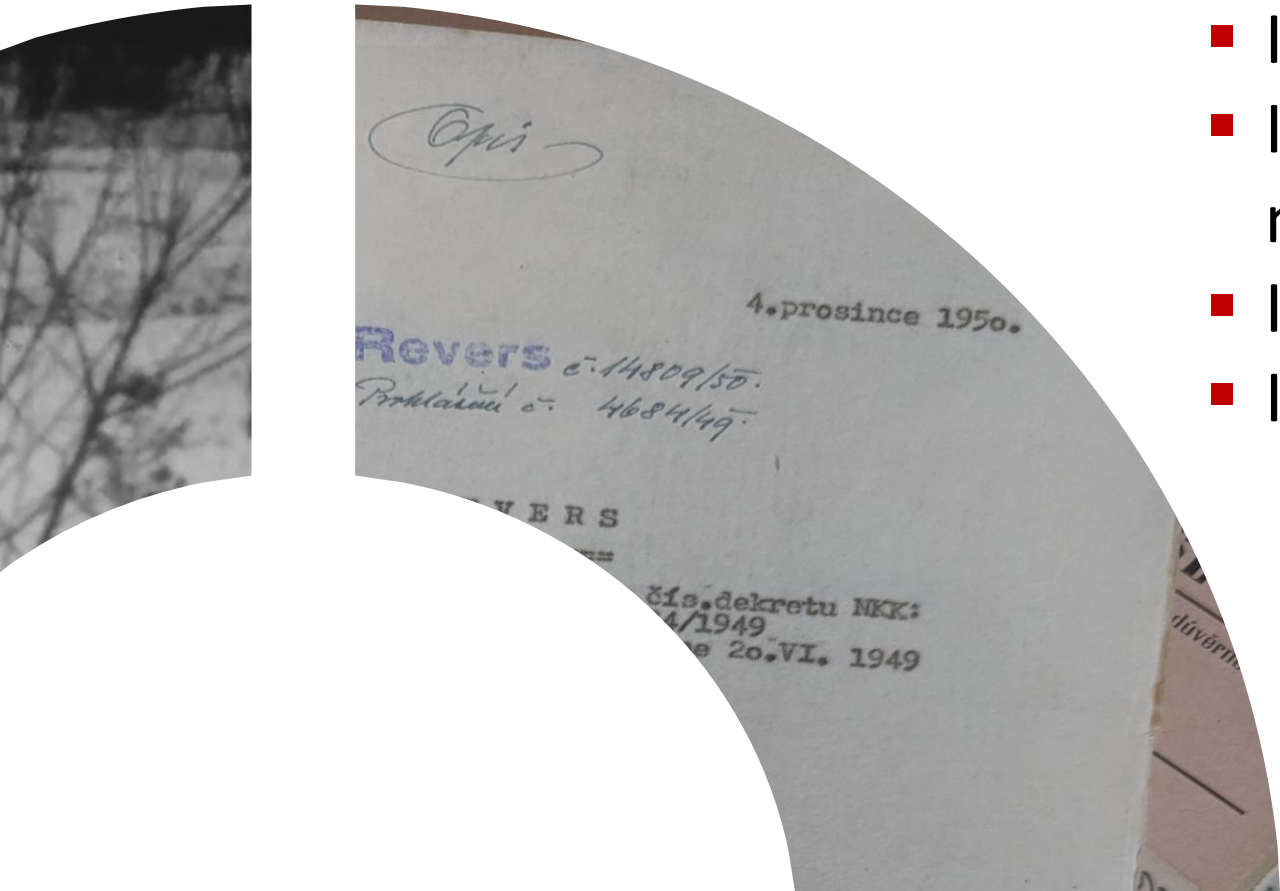
# Co to znamená?

- 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**



# Scénář I. – výběr dle seznamu pro NArP

IV č. 85/2024 (část II)

## Oznámení Ministerstva vnitra, kterým pro elektronické systémy

Ministerstvo vnitra zveřejňuje na základě § 70 odst. 1 písm. b) zákona č. 106/1999 Sb., o archivnictví a spisové službě a o změně některých zákonů (dále jen „zákon“), národní standard pro elektronické systémy (dále jen „národní standard“).

Předkládané znění národního standardu vychází z předchozího znění, především reaguje na zákon č. 197/2024 Sb., kterým se mění zákon č. 106/1999 Sb., o archivnictví a spisové službě a o změně některých zákonů, ve znění zákonů č. 261/2021 Sb., kterým se mění některé zákony v souvislosti s přijetím elektronizací postupů orgánů veřejné moci, ve znění pozdějších předpisů, a zákon č. 106/1999 Sb., o svobodném přístupu k informacím, ve znění pozdějších předpisů, a vyhláškou č. 200/2024 Sb., kterou se mění vyhláška č. 259/2010 Sb., o provádění spisové služby, ve znění pozdějších předpisů a dále oprávněné zjevné chyby předchozího znění zveřejněného ve Věstníku Ministerstva vnitra č. 24, 2024.

zvýšenou kursivou není integrální součástí požadavků samýci  
světlení smyslu požadavku nebo poskytnutí ilustračních  
v praxi.

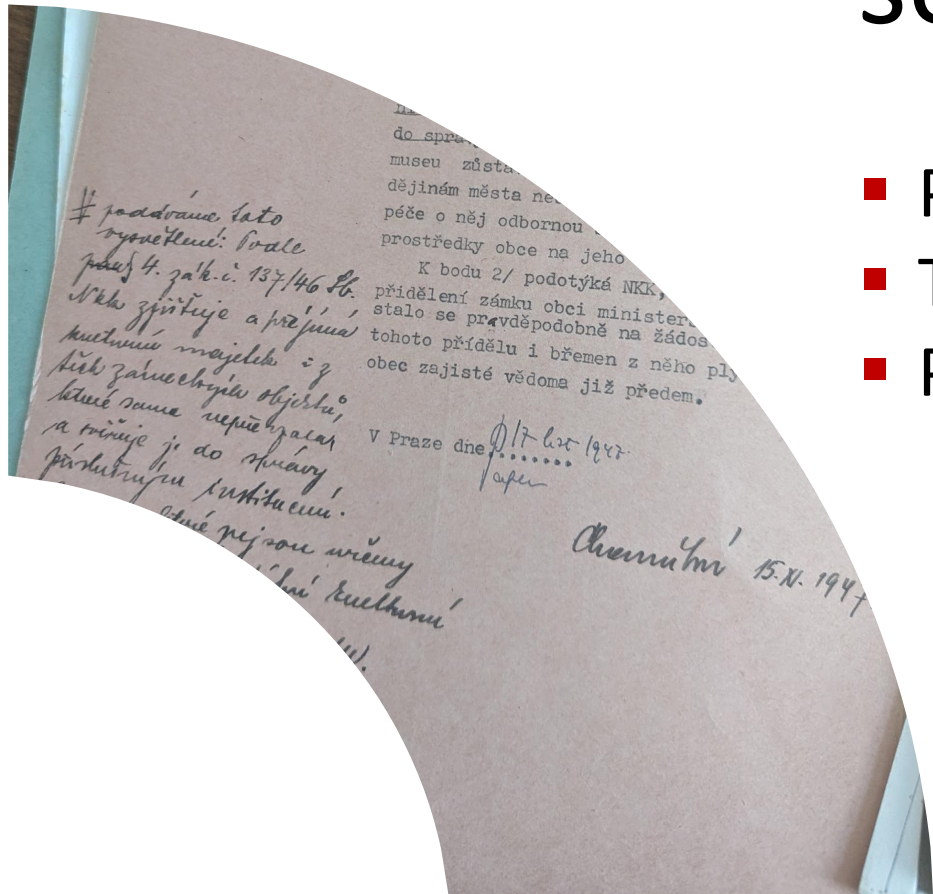
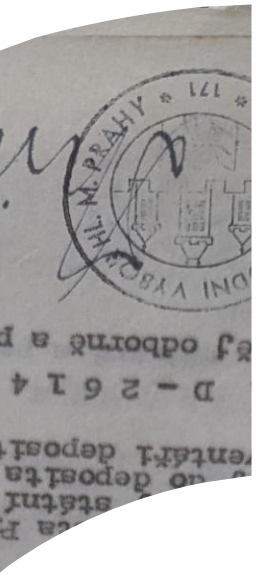
standardu:

měna textu
stranění odkazu na přílohu č. 3
na textu
1
xtu
tu
tu
"

- 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



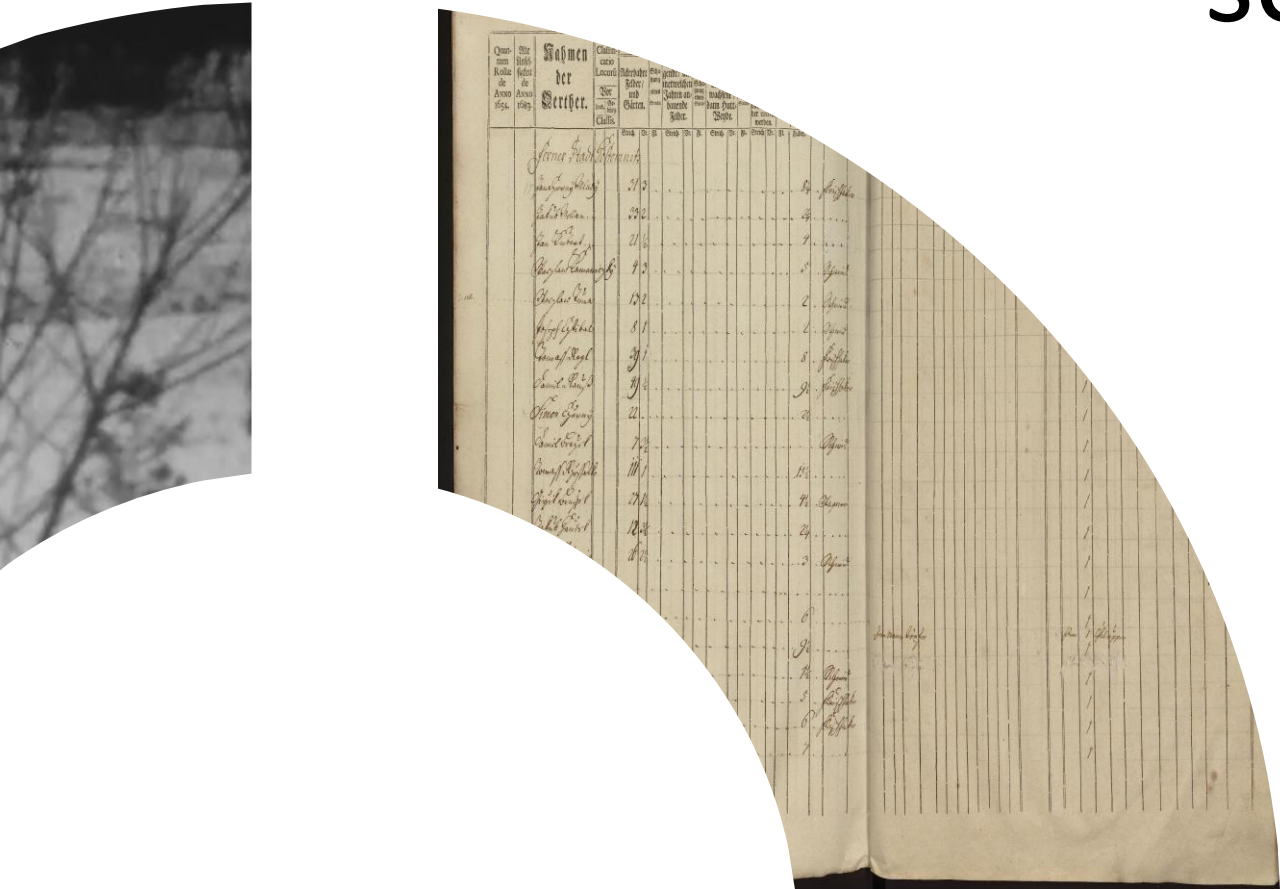


# Scénář I. – výběr dle seznamu pro NArP

- 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



Czech OCR Archivace

Vstupní soubor:  [Browse](#)

Nebo složka:  [Browse](#)

Výstupní CSV:  [Save As...](#)

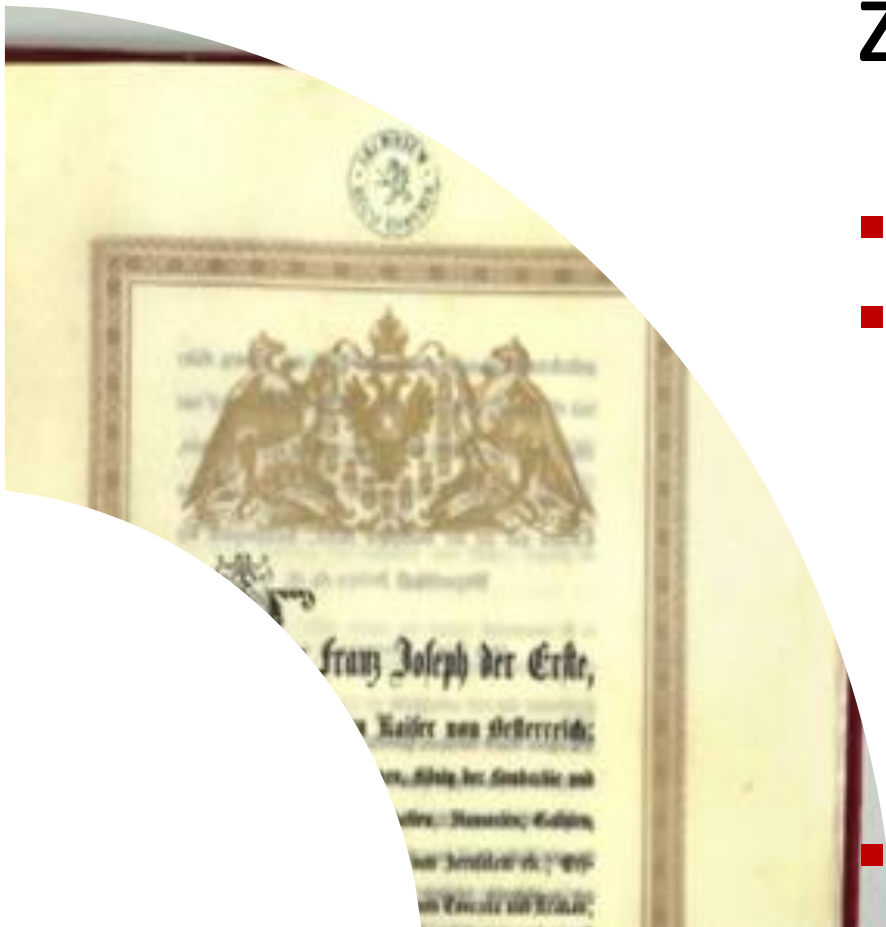
Váhy (volitelné JSON):  [Browse](#)

Typ instituce (výběr uživatele):  [▼](#)

[Spustit](#) [Ukončit](#)

# 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



newcontentonline.lovable.app

**Nahrát Dokument**  
Vyberte soubory nebo přetáhněte

Camera Files

**Kontext Dokumentu**

Účel Archivace **Povinné**

Státní Archiv Osobní Archiv Firemní Archiv

Typ Původce **Povinné**  
Vyberte typ původce

Časové Rozmezí **Povinné**  
např., 2020-2025

Unikátnost **Povinné**  
Vyberte úroveň unikátnosti

Související Osoby (Volitelné)  
Jména oddělená čárkami

Legislativa (Volitelné)  
Související zákony nebo nařízení

Poznámka (Volitelné)  
Dodatečné poznámky nebo postřehy

Analyzovat Dokument

**DocuMate**  
AI rozhodovací engine

Originál

Související Osoby (Volitelné)  
Jména oddělená čárkami

Legislativa (Volitelné)  
Související zákony nebo nařízení

Poznámka (Volitelné)  
Dodatečné poznámky nebo postřehy

**Vyřadit Po Skartační Lhůtě**  
Schedule for disposal from year 2031.

Skartovat od roku  
**2031**

Jistota: 9/10  
Edit with Lovable



# Scénář IV. – DocuMate App

## ■ Testování kritérií a parametrů

**Rozhodovací Matice**  
Spravovat CSV soubory

Rozhodovací Matice Váhy Kritérií Podrobná kritéria

### Nahrát Rozhodovací Matici

Nahrajte CSV soubory obsahující rozhodovací pravidla pro klasifikaci archivace dokumentů

Účel Archivace  
Státní Archiv

Typ Původce  
Obce I. typu

CSV Soubor  
Vybrat soubor Soubor nevybrán Nahrát

Maximální velikost souboru: 5MB. Pouze .csv soubory.

### Existující Rozhodovací Matice

Předpřipravená rozhodovací matice pro "obce I. typu"

decision-matrix-obce-I-typu.csv  
220 typů dokumentů • Czech language Stáhnout

**Rozhodovací Matice**  
Spravovat CSV soubory

Rozhodovací Matice Váhy Kritérií Podrobná kritéria

### Váhy Kritérií

Nastavte váhy pro jednotlivá kritéria. Součet všech vah musí být 1000 bodů. Exportovat Váhy

Účel Archivace	200 bodů
Typ Původce	200 bodů
Časové Rozmezí	100 bodů
Unikátnost	150 bodů
Legislativa (Volitelné)	100 bodů
Popis Obsahu	100 bodů
context.purposeAndUsage	50 bodů
context.documentDirection	50 bodů
Archivní Série (Volitelné)	50 bodů

**Rozhodovací Matice**  
Spravovat CSV soubory

Rozhodovací Matice Váhy Kritérií Podrobná kritéria

### Významné osoby

Seznam významných osob, které budou použity při vyhodnocení souvisejících osob

Soubor (CSV nebo TXT)  
Vybrat soubor Soubor nevybrán Nahrát

Příklad formátu:  
Formát: Jméno Příjmení, Funkce/Význam  
Jan Novák, Ministr kultury  
Marie Svobodová, Historička

### Legislativa

Seznam relevantní legislativy a právních předpisů

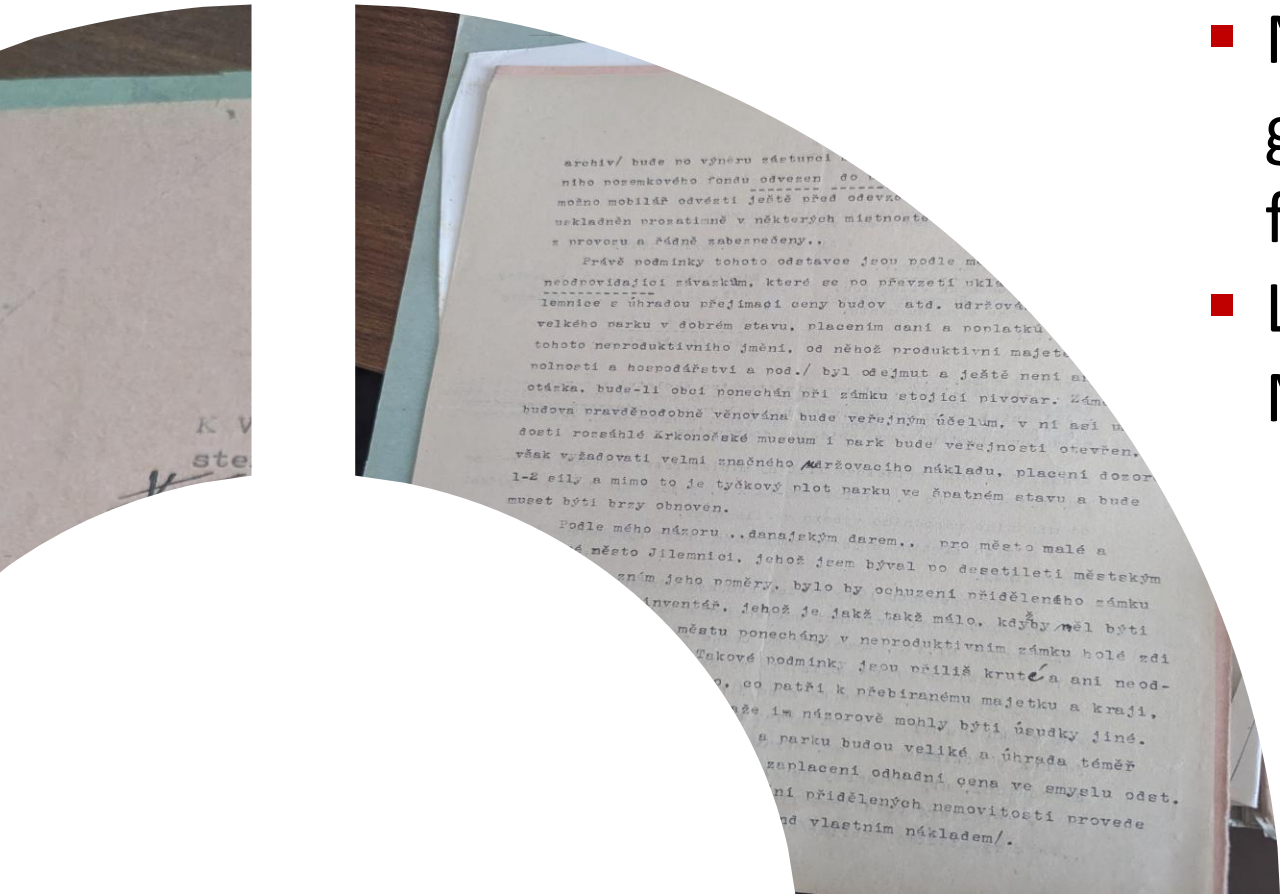
Soubor (CSV nebo TXT)  
Vybrat soubor Soubor nevybrán Nahrát

Příklad formátu:  
Formát: Číslo zákona, Název  
499/2004 Sb., Zákon o archivnictví  
181/2000 Sb., Zákon o ochraně osobních údajů

### Události

Seznam významných událostí relevantních pro archivování





# 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

Pavλίna 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

Návrh vyhlášky o vydání zlaté mince "Městská památková rezervace" KČ

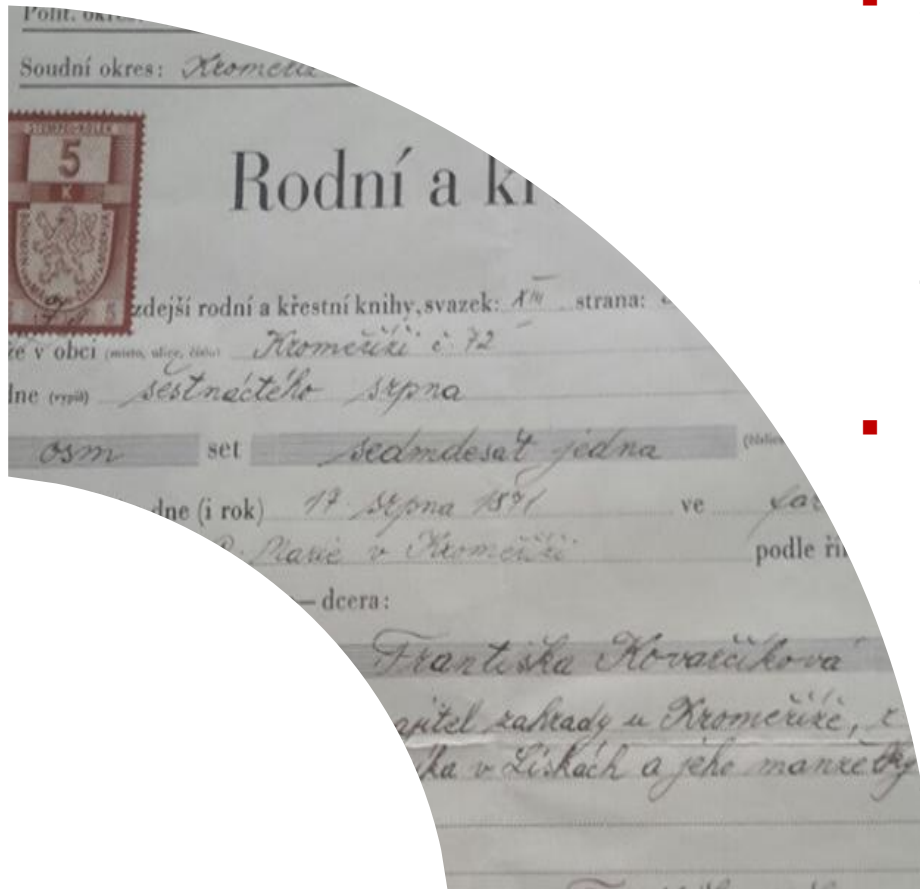
Přehled Přípo.

Identifikace materiálu		Zpracování	
Čj. OVA		Stav materiálu	
Čj. předkladatele	2025/045942/CNB/001	Datum schůze	
PID	KORNDG7C2TOT	Leg. proces pokračování	
Autorizace	29. dubna 2025 10:40	Číslo sněmovního tisku	
Správa	19. září 2025 12:31	Číslo senátního tisku	

Návrh vyhlášky o vydání zlaté mince "Městská památková rezervace" KČ	Návrh vyhlášky
Česká národní banka	bankovníctví
Finanční právo/Bankovníctví a pojišťovnictví	Zákon č. 6/1993 Sb.
Návrh vyhlášky o vydání zlaté mince "Městská památková rezervace" KČ	

# 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

The screenshot shows a web application interface for material identification. The main title is "Návrh vyhlášky o vydání zlaté mince "Městská památková rezervace KČ". Below the title, there are buttons for "Přehled" and "Připoj.". The interface is divided into two main sections: "Identifikace materiálu" and "Zpracování".

Identifikace materiálu	
Čj. OVA	
Čj. předkladatele	2025/045942/CNB/001
PID	KORNDG7C2TOT
Datum autorizace	29. dubna 2025 10:40
Datum správy	19. září 2025 12:31

Zpracování	
Stav materiálu	
Datum schůzky	
Leg. proces pokračování	
Číslo sněmovního tisku	
Číslo senátního tisku	

Návrh vyhlášky o vydání zlaté mince "Městská památková rezervace KČ"
Návrh vyhlášky
Česká národní banka
bankovníctví
Finanční právo/Bankovníctví a pojišťovnictví
Zákon č. 6/1993 Sb.
Návrh vyhlášky o vydání zlaté mince "Městská památková rezervace KČ"