

The logo consists of a black circle on a yellow background. Inside the circle, the text "Národní filmový archiv" is written in a bold, black, sans-serif font, arranged in three lines.

**Národní
filmový
archiv**

Chatbot, který rozumí sbírce?

Lokální LLMs architektura RAG v
paměťových institucích

Jonáš Kucharský, Národní filmový archiv
AKM 2025

Motivace a rámec

- **Velké sbírky, vysoká heterogenita, historické nánosy.**
 - katalog vs. partikulární inventární data
- **Nedostatek jednotných slovníků/terminologie.**
 - identifikace fondů v polních podmínkách,
 - interpretace produkčních metadat
- **Nutnost zpracování velkého objemu informací**
 - projekty masové digitalizace (NPO)
 - identifikace titulů
 - priority digitalizace
- **3 případové studie:**
 - 1) čištění dat a tvorba řízeného slovníku
 - 2) scraping dat a analýza kanonického postavení děl
 - 3) konverzační RAG systém

	+		H	I	K	L	Q	S	T	U
	A	F								
1	ID	titul - krabice (film)	formát	typ materiálu - link	zvuk / sound typ záznamu / recording type (boční strar	díl / part	ozvučen dne / c	středisko / studio	úsek	
	p-100000	Svejk VI	35 mm ▾	207 LN ▾	Ing. Detak ▾	mez. mix - Kopie	1. + 2. díl	14.1.87	3242	C.T. 424517
3	p-100001	inten. left blank	- ▾	- ▾	- ▾	-	-	-	-	-
4	p-100002	Růžový slon	35 mm ▾	205 LN ▾	Mader ▾	přepis mix (nový)	-	18.10.1989	3242	č.f. 428802
5	p-100003	Čardášová Julie	35 mm ▾	206 LN ▾	- ▾	mez. pás/ mezmix	1. + 2. díl	20.12.85	-	-
6	p-100004	Hřebínek	35 mm ▾	205 LN ▾	Ing. Děták ▾	nový mezmix	-	15.10.87	3242	-
7	p-100005	Pražské mosty	35 mm ▾	206 LN ▾	- ▾	mez. pás, hudba, ruchy	1. díl	26.06.84	-	-
8	p-100006	Krušnohorské pohádky	35 mm ▾	205 LN ▾	Maden ▾	mix	1. díl	8.9.89	3242	č.f. 427523
9	p-100007	Žofka 13	35 mm ▾	205 LN ▾	Ing. Děták ▾	komentář německy	-	19.06.1989	3241	č.f. 419901
10	p-100008	Šťastný swing	35 mm ▾	206 LN ▾	- ▾	mez. pás	1. + 2. díl	20.12.85	-	-
11	p-100009	Psí Kusy	35 mm ▾	205 LN ▾	- ▾	mez. pás, hudba, ruchy (mez mix)	4. díl	03.10.89	-	-
12	p-100010	Unikát	35 mm ▾	206 LN ▾	- ▾	mez. pás	1. díl	18.12.84	-	-
13	p-100011	Poetická cvičení	35 mm ▾	206 LN ▾	Ing. Děták ▾	mez. mix (mez pás)	-	26.09.86	3242	č.f. 426521
14	p-100012	Myšlenka	35 mm ▾	206 LN ▾	- ▾	mezinár. pás	-	20.3.	-	-
15	p-100013	Psí Kusy	35 mm ▾	205 LN ▾	- ▾	mezinárodní pás, hudba, ruchy (mez. mix)	5. + 6. díl	23.11.1990	-	-
16	p-100014	Co vy na to, pane barone?	35 mm ▾	205 LN ▾	- ▾	mez. pás, hudba, ruchy	1. díl	26.06.84	-	-
17	p-100015	Karel IV	35 mm ▾	205 LN ▾	Ing. Děták ▾	mezmix	-	20.12.1988	3242	č.f. 427534
18	p-100016	Pat a Mat	35 mm ▾	207 LN ▾	- ▾	mix=mez. pás	6. díl	03.12.1990	-	-
19	p-100017	O Princezně Furieně	35 mm ▾	206 LN ▾	- ▾	mix španělsky	1. díl	05.05.1987	3242	-
20	p-100018	O Hloupém peciválovi	35 mm ▾	205 LN ▾	- ▾	mezinárodní pás (mezmix)	1. + 2. díl	06.05.89	3242	-
21	p-100019	Rohy	35 mm ▾	205 LN ▾	Ing. Děták ▾	mix portugalská v.	1. + 2. díl	19.8.1987	3242	-
22	p-100020	Návrat zlaté rybky	35 mm ▾	205 LN ▾	Ing. Děták ▾	mix=mezmix	-	25.3.1988	3242	č.f. 427517
23	p-100021	Eberhard	35 mm ▾	206 LN ▾	- ▾	mez. pás, sbory	3. + 4. díl	29.11.84	-	-

Proč LLMs?

- **Umí pracovat s nečistými, historicky vrstvenými daty**
 - varianty názvů, nejednotné popisy, šumy v terminologii
 - LLM zvládne navrhnout normalizaci nebo porovnání (fuzzy shoda + vysvětlení)
- **Umožňují generovat návrhy, ne automatická rozhodnutí**
 - ideální pro human-in-the-loop workflow
 - archivář vždy rozhoduje → LLM jen asistuje
- **Urychlují opakované kurátorské úlohy**
 - čištění hodnot, mapování polí, návrh kontrolovaných slovníků
 - tvorba shrnutí, klasifikace, verifikace konzistence
- **Otevírají nové způsoby přístupu ke sbírkám**
 - konverzační vyhledávání (RAG)
- **Nejde o „umělou inteligenci“, ale o nástroj**
 - využitelný tam, kde pravidla selhávají nebo jsou příliš nákladná (nedostatek lidí a času)
 - flexibilní, opakovatelný, přizpůsobitelný doméně

LLMs v paměťových institucích: rizika a požadavky

- **LLM ≠ plug-and-play**
- **Ochrana dat a suverenita**
 - omezení předávání dat do cloudu
 - lokální modely na vlastním hardwareu (Ollama) vs. cloud based služby
- **Open weights ≠ open source**
- **Auditovatelnost procesů**
 - nutnost sledovat, z čeho odpověď vznikla
 - dohledatelnost dotazu, výsledků a chování modelu
- **Minimalizace halucinací**
 - LLM jako asistent
 - human-in-the-loop při kritických operacích
- **Kombinace pravidel + LLM (LLM nenahrazují kurátora).**

Případová studie 1: čištění dat, řízený slovník

Výchozí situace:

- mg pásy zpracované v NFA, na 5000 záznamů
- stovky varianty názvů titulů, 517 variant zápisu typu záznamu

Cíl:

- sjednotit terminologii
- navrhnout jednoznačnou ontologii typů zvukových záznamů

Navržená ontologie:

- Vytvořena v dialogu kurátor + LLM
 - Technický typ (mix, mezinárodní pás, přepis, kopie, synchron, ...)
 - Obsahový typ (hudba, ruchy, dialogy, komentář, zpěv, hlasy...)
 - Sekundární charakteristiky (stereo/mono, verze, zkouška, nový/starý, ...)
 - Jazyk (cs, de, en, ru...)

PS1: LLM jako auditor konzistence

- Samostatný LLM skript porovnal výsledky pravidel s vlastní „interpretací“.
- Vrátil JSON: správnost, navržené opravy, jistota.
- Kurátor měl přehled o sporných případech, kontrola zůstala lidská.

Výsledek:

- Konsolidovaná, bilingvní tabulka zvukových materiálů.
- Jasně identifikovatelné zvukové zdroje ke každému filmu.
- Kvalitní podklad pro určování priorit digitalizace a plánování péče o audiofondy a jejich budoucího zpřístupňování

Použité nástroje

- Python, Pandas, heuristiky pro normalizaci, regex, LLM (OpenAI / Gemini / lokál LLMs) jako asistent při návrhu pravidel, LLM jako auditor konzistence

1	original_term	count	normalized_term	technical_type_c	secondary_char	content_types_cs	has_multiple_content	language	needs_review
403	mix francouzská v.	1	Mix francouzská verze	mix	verze	[neurčeno]	no	fr	maybe - technical
404	Mez.mix Hudba	3	Mez. mix Hudba	mix		hudba	no		no
405	mez mix - hudba	3	Mez. mix - hudba	mix		hudba	no		no
406	mix přepis/hudba	1	Mix přepis/hudba	mix		hudba	no		no
407	hudba mix česky	1	Hudba mix česky	mix		hudba	no	cs	no
408	Hudba - Mez.mix	1	Hudba - Mez. mix	mix		hudba	no		no
409	mezinárodní mix	4	Mezinárodní mix	mix		hudba; ruchy	yes		no
410	mezinár. mix	3	Mezinár. mix	mix		hudba; ruchy	yes		no
411	mix mezinár. pás	2	Mix mezinár. pás	mix		hudba; ruchy	yes		no
412	mix mezinárodní pás	1	Mix mezinárodní pás	mix		hudba; ruchy	yes		no
413	mezinárodní pás=mix	1	Mezinárodní pás=mix	mix		hudba; ruchy	yes		no
414	mezinárodní pás (= mix)	1	Mezinárodní pás (= mix)	mix		hudba; ruchy	yes		no
415	mez. pás, hudba, ruchy (mez mix)	1	Mez. pás, hudba, ruchy (Mez. r	mix		hudba; ruchy	yes		no
416	mezinár. mix cop	1	Mezinár. mix cop	mix		hudba; ruchy	yes		no
417	mezinárodní pás mix německá ver.	1	Mezinárodní pás mix německá	mix		hudba; ruchy	yes	de	no
418	mezinárodní pás mix německy	1	Mezinárodní pás mix německy	mix		hudba; ruchy	yes	de	no
419	mezinár. mix (cop)	1	Mezinár. mix (cop)	mix		hudba; ruchy	yes		no
420	mezinárodní pás hudba +zpěv přepi	1	Mezinárodní pás hudba +zpěv	mix		hudba; zpěv/sbory	yes	cs	no
421	Mix - bez komentáře - Mez. pás	1	Mix - bez komentáře - Mez. pás	mix		komentář	no		no
422	mix efekty	9	Mix efekty	mix		ruchy	no		no
423	Mix efekty	5	Mix efekty	mix		ruchy	no		no
424	Mez mix - ruchy	4	Mez. mix - ruchy	mix		ruchy	no		no
425	Mez.mix Ruchy	4	Mez. mix Ruchy	mix		ruchy	no		no
426	formix ruchů	1	Formix ruchů	mix		ruchy	no		no
427	Mix mez pasy ruchy	1	Mix mez pasy ruchy	mix		ruchy	no		no
428	znělka mix	1	Znělka mix	mix		znělka	no		no
429	přemix	8	Přemix	přemix		[neurčeno]	no		maybe - technical
430	premix	1	Premix	přemix		[neurčeno]	no		maybe - technical
431	přemix (mix přepis)	1	Přemix (mix přepis)	přemix		[neurčeno]	no		maybe - technical

Případová studie 2: scraping a obohacování dat

Výchozí problém: Inventární CSV obsahovaly nejednoznačné zápisy (přítomny na samotných nosičích)

- varianty názvů, překlepy, nejednotné formáty.
- to komplikuje párování s katalogem, identifikaci a rozhodování o prioritách digitalizace

Cíl mini-projektu: Vyčistit a zpřesnit inventární data, obohatit data o vkusová „soft metadata“

- LLMs
 - identifikace chybných či nejednoznačné názvů, sjednocení variant, ověření shody mezi zdroji (db NFA, wikidata, veřejné filmové databáze)
 - analýza soft medatat
 - hodnocení na CSFD,
 - analýza článků publikovaných na portálu Filmový přehled,
 - analýza interního hodnocení NFA)

PS2: obohacování dat

Co jsme doplnili nad rámec základních názvů:

- Crowdsourcované ratingy z ČSFD (jednoduchý signál zájmu/známosti).
- Anotace, ocenění a články z Filmového přehledu / Revue (odborný kontext).
- LLM zpracovala obě báze pro vyhodnocení vnímaného kanonického postavení díla, s jasným linkem na zdroj

Použité nástroje:

- ad-hoc skripty pro scraping databází a článků
 - Python, Selenium, BeautifulSoup, SerpAPI
- LLM služby
 - OpenAI GPT-4o, Google Gemini 2.0, Ollama (Gemma, LLaMA, Mistral aj.)

	A	B	E	F	G	H	I	J	K	L	M	N	O	
1	ais_id	Název filmu	year	director	csfd_rating	csfd_1_5	gemini_rating	new_gpt_rating	AIS_RATING_AVERAGE	g_music_score	g_note	g_music	source	
2	0011179	Vlak do stanice nebe	1972	Kachyňa Karel	73	2	2	1	3.5	3	Despite numerous a	Although Jiří Lenoch	Filmový přehled data	
3	0011948	Srdečný pozdrav ze	1982	Lipský Oldřich	67	2	2	2	3.2	3	A quirky sci-fi come	The film received no	Filmový přehled data	
4	0011339	Discopřiběh	1987	Soukup Jaroslav	52	3	3	4	2.5	3	A typical populist fil	The film's annotatio	Filmový přehled data	
5	4000513	Bajaja	1950	Trnka Jiří	74	2	2	1	0.3		2	A melancholic pupp	The film won a Natio	Filmový přehled data
6	0011805	Od vraždy krok ke lži	1982	Tuček Petr	44	3	3	3			3	Without annotation,	The music score ber	Filmový přehled data
7	0011341	Jonáš a Melicharová	1986	Sis Vladimír	41	3	3	4			3	A nostalgic patchwo	No local awards for	Filmový přehled data
8	0011832	Papilio	1986	Svoboda Jiří (režie)	65	2	3	2	3.3	3	A war psychological	Although Jozef Reva	Filmový přehled data	
9	0011160	Báječní muži s klikou	1978	Menzel Jiří	71	2	2	1	4.1	3	A nostalgic and play	The annotation highl	Filmový přehled data	
10	0011344	Lev s bílou hřívou	1986	Jireš Jaromil	63	2	3	2	3.8	3	A poetically-driven b	The film won an awa	Filmový přehled data	
11	0011400	Prodáváč humoru	1984	Krejčík Jiří	64	2	3	2	1.6	3	A satiric comedy crit	The music was mad	Filmový přehled data	
12	0011145	Adéla ještě nevečeře	1977	Lipský Oldřich	61	2	2	1	4.8	3	A clever and humor	While Jiří Lenoch is	Filmový přehled data	
13	0012031	Antonín Dvořák	1991	Jireš Jaromil	73	2	4	3	3.0	3	Based on the limited	Given the lack of aw	Filmový přehled data	
14	0011249	Jak se budí princezn	1977	Vorlíček Václav	78	2	2	2		2	A well-received and	Karel Svoboda's mu	Filmový přehled data	
15	0011668	Přiliš velká šance	1984	Fuka Otakar	65	2	3	3		3	A moderately engag	The music and soun	Filmový přehled data	
16	0010339	Něffňukej, veverko!	1988	Plívová-Šimková Vě	54	3	2	1	3.9	3	A successful childre	While Antonín Holub	Filmový přehled data	
17	0010350	Uzavřený okruh	1989	Matějka Václav	62	2	3	3	1.9	3	A critically-minded	The available data d	Filmový přehled data	
18	0010372	Masseba (Pohled z c	1989	Zábranský Miloš	57	3	3	3	1.2	3	Without further infor	Given the absence o	Filmový přehled data	
19	0011361	Šašek a královna	1987	Chytilová Věra	66	2	2	1	4.1	2	A formally experime	Jiří Bulis won the aw	Filmový přehled data	
20	0010376	Přiběh 88	1989	Hojdová Zuzana	55	3	4	3	1.5	3	Lacking substantial	Based on the availa	Filmový přehled data	
21	0011126	Hodina modrých slon	1971	Cvrček Radim	57	3	3	2		3	A children's film wit	While no explicit crit	Filmový přehled data	
22	0010589	Sedmé nebe	1987	Kosek Otakar	52	3	3	3		3	A mildly humorous t	Petrov Vadim and K	Filmový přehled data	
23	0011274	Do zubů a do srdíčka	1985	Drha Vladimír	51	3	3	2	2.4	3	A sympathetic, thou	No local data suggest	Filmový přehled data	
24	0011020	Petrolejové lampy	1971	Herz Juraj	84	1	2	1	4.9	3	An expressively dire	While the film receiv	Filmový přehled data	
25	0011352	Pavučina	1986	Zaoral Zdenek	63	2	3	2	4.0	3	Pavučina' appears to	Though Václav Hálek	Filmový přehled data	
26	0011370	Mezi námi kluky	1981	Cvrček Radim	42	3	3	2	1.3	2	A lighthearted come	The film won a "Cen	Filmový přehled data	
27	0012086	Začátek dlouhého p	1990	Hledík Peter	52	3	4	3	1.3	3	The film, despite bei	With music by Petr	Filmový přehled data	
28	4001597	Pohádky tisíce a jed	1974	Zeman Karel	85	1	2	1	3.6	3	The film likely exhib	Given the absence o	Filmový přehled data	
29	0011116	Na kometě	1970	Zeman Karel	83	1	2	1		3	A visually inventive b	The film does not ap	Filmový přehled data	
30	0011995	Únos Moravanky	1982	Muchna Milan	32	4	4	4		3	A lighthearted folk c	The film's reliance o	Filmový přehled data	

PS3: RAG

Retrieval-Augmented Generation (RAG):

- architektura kombinující:
 - RETRIEVAL – vyhledání relevantních dokumentů v databázi (např. v Elasticsearch, katalogu, textovém korpusu). → Do LLM vstupují pouze reálná data.
 - GENERATION – jazykový model na základě těchto dokumentů vytvoří odpověď ve srozumitelném jazyce. → LLM negeneruje ze svých tréninkových dat, ale vysvětluje to, co našel.
- **RAG minimalizuje halucinace**
 - LLM má odpovídat z dat, co mu dáme.
 - odpovědi jsou založené na datech.
 - umožňuje konverzační vyhledávání nad archivními sbírkami bez nutnosti měnit jejich strukturu.

Lokální LLMs

- **TinyLlama (1–2B)** – běží i na CPU, minimální nároky.
- **Mistral 7B** – efektivní malý model, stačí 8–12 GB VRAM.
- **Llama 2 / Vicuna 13B** – vyvážený výkon, ideál pro 16–24 GB VRAM.
- **Qwen 2/3 (7–14B)** – moderní, velmi kvalitní, nároky podobné Llama 13B.
- **Qwen 3 72B / Llama 3 70B** – SOTA varianta pro více-GPU servery.

PS3: RAG v archivním prostředí

Problém, který RAG řeší:

- Uživatelé i kurátoři potřebují rychlé kontextové odpovědi, ne listy výsledků
- Data jsou rozprostřená v různých polích, tabulkách, materiálech.
- U starších záznamů chybí jednotná terminologie a přímé odkazy.

Co RAG přináší:

- Konverzační přístup ke sbírce: ptám se stejně, jako bych se ptal archiváře.
- Odpovědi založené na zdrojových datech, ne na halucinacích.
- Možnost postupných dotazů („A kdo to režíroval?“ → „A máme digitální kopii?“).

Výhody pro instituci:

- Lze provozovat lokálně (Ollama + Elasticsearch).
- Auditovatelné dotazy i odpovědi, úspora času, rychlejší orientace ve fondech.

PS3: experimentální RAG NFA

Architektura (zjednodušeně):

- **Datová vrstva (Elasticsearch)** – index katalogu filmů, strukturovaná metadata, materiály, autoři.
- **MCP (Model Context Protocol)**
 - jednotná brána mezi RAG a ES: náš kód nikdy nevolá ES přímo, vždy přes MCP adaptér
- **Konfigurační modul**
 - dodá znalost struktury ES (mapping, nested vs. simple), specifikuje logování dotazů, bezpečnost
- **Query Generator**
 - mozek nástroje, analýza konverzačního kontextu
 - detekce režimu dotazování
 - extrakce klíčových polí,
 - volba odpovědního modu
 - generování vhodného ES dotazu
- **Answer Generator**
 - zpracování výstupu z ES query
- **Conversation manager, logging**
 - konverzační paměť, auditovatelnost, analýza

PS3: jak systém pracuje

1) Analýza dotazu

LLM + heuristiky určují, *o co se uživatel ptá* (film, osoba, materiál, digitální kopie, specifický filtr...).

2) Volba režimu dotazování (módy)

- *pattern* – pevná pravidla (rychlé, spolehlivé pro jednoduché dotazy),
- *llm-enhanced* – kombinace pravidel a LLM,
- *pure-llm / reinforcement* – pokročilé módy.

3) Generování ES dotazu

QueryGenerator využije znalost schématu (nested, analyzéry) a vybuduje validní Elasticsearch DSL.

4) Získání a příprava dokumentů

DocumentPreprocessor z metadat vytvoří stručné texty vhodné pro LLM.

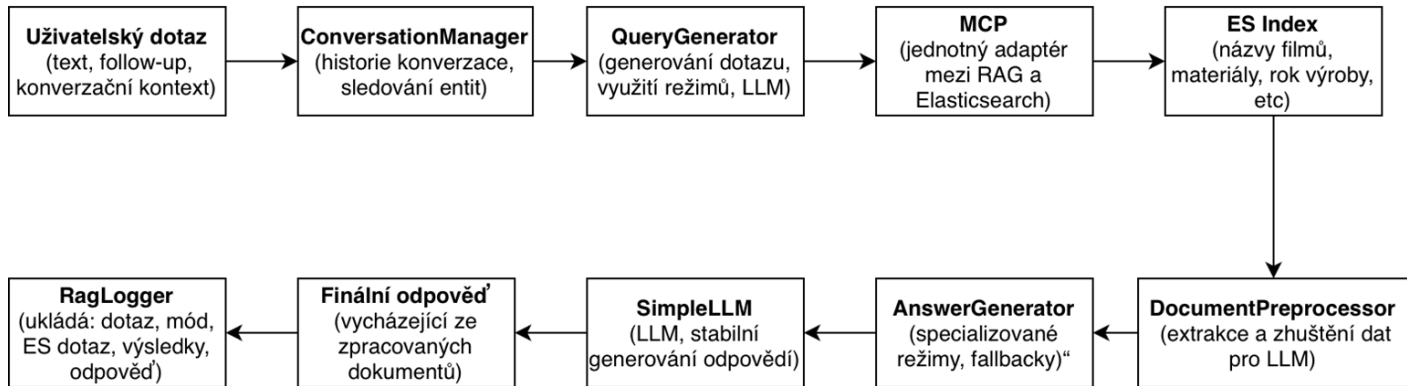
5) Generování odpovědi

AnswerGenerator vytvoří srozumitelný výstup: seznam filmů, shrnutí, filtry, vysvětlení.

6) Konverzační paměť

Umožňuje follow-up otázky („a které mají HD kopii?“) a auditovatelnost.

Schéma RAG systému



PS3: limity systému?

Silné stránky:

- jednotná vrstva MCP → čisté API,
- bezpečné oddělení RAG od Elasticsearch.
- lokální LLMs, nehrozí úniky dat
- kombinace pravidel + LLM → přesné dotazy i při nejednoznačných otázkách.
- konfigurace → snadná změna modelu, indexu, režimů bez zásahů do kódu.
- auditovatelnost → vše se loguje, lze rekonstruovat celý postup.

Slabiny:

- lokální LLMs :)
- bez vektorového vyhledávání → horší práce s volným jazykem a obsahovými podobnostmi.
- pattern heuristiky → vyžadují údržbu, nejsou škálovatelné na nové typy dat.
- LLM generuje ES dotazy ad-hoc → někdy vede k překomplexním DSL nebo menší stabilitě .
- konverzační vrstva je lokální → paměť je jednoduchá, nepracuje s dlouhodobým kontextem.

Možné alternativy

- hybridní vyhledávání (vektorový index + ES)
- agentní systém
- post training :)