

Technická specifikace

Stručný popis výzkumného záměru

Cílem projektu je vytvoření a implementace *workflow* a softwarových nástrojů pro automatické generování popisných metadat pro digitalizáty agregované v Manuscriptoriu (středověké rukopisy a jiné dokumenty písemného a dokumentového dědictví) pomocí umělé inteligence (AI) a dostupných aplikačních rozhraní k AI. Tato nově vytvořená klíčová slova a jiné popisné informace poslouží k obohacení obsahu Manuscriptoria, což povede ke zlepšení třídění a vyhledávání těchto historických dokumentů, jakož i k implementaci nových funkcí v koncovém uživatelském rozhraní zkvalitňujících uživatelský komfort.

Pozadí a smysl záměru

Manuscriptorium shromažďuje obsah ze stovek zdrojů. Jedná se o obrazy i popisná metadata. Existence obrazů je pro uživatele zásadní, ale kvalita popisných dat je neméně důležitá, protože jednak popisy pomáhají orientaci v digitalizovaném obsahu a zároveň pomáhají porozumět informačnímu obsahu v obrazech zachycenému. Existence kvalitních strukturovaných popisů navíc nejenom zásadně usnadňuje vyhledání digitalizátů, ale také je významným heuristickým prostředkem pro objevování nových témat a formulování netradičních problémů.

Pokud jde o agregaci a popisná metadata, pak tato jsou partnery poskytována v různé kvalitě, ať již jde o informační rozsah, strukturu nebo využívání referencí na položky formálních slovníků. Manuscriptorium si nikdy nekladlo za cíl klasifikovat (známkovat) kvalitu partnery poskytovaných dat a metadat (s výjimkou všeobecné osvěty nebo explicitně vyžádané podpory pro dosažení *best practice* při plnění role Manuscriptoria jako doménového agregátora pro Europeana).

Manuscriptorium (Národní knihovna České republiky) se úspěšně zapojila do projektu ARMA, ve kterém vznikla *best practice* pro optimální EDM záznam pro středověké rukopisy (*EDM ARMA Specification*). Pro naplnění setu ARMA v Europeana a splnění závazků v tomto projektu bylo nutno ručně do agregovaných záznamů doplňovat chybějící informace, aby výsledné záznamy odpovídaly nově definované *best practice*. Bylo sice nastaveno potřebné *workflow* tak, aby tyto nové informace mohly koexistovat s originálním obsahem, přirozeně se však potvrdilo, že pro systematickou práci tohoto typu v měřítku Manuscriptoria (statisíce popisných metadat) chybějí (a vždy chybět budou) lidské zdroje.

Naším cílem je tedy prozkoumat potenciál využití generického modelu dostupných AI technologií a prověřit, zda jsme schopni získat dostatečně kvalitní metadata s minimálními zdroji (čas, finance) a to bez nutnosti trénování vlastních speciálních modelů (kvůli minimalizaci nákladů na zdroje).

Předběžné analýzy provedené na jaře 2024 v součinnosti Objednatele s Poskytovatelem naznačují již nyní vynikající poměr cena zdrojů vs. kvalita metadat při použití ChatGPT 40. Využívání API je sice zpoplatněno, ale výsledky získané s využitím generických modelů vypadají slibně.

Očekávané výsledky

- Efektivní plynulé pořizování přesných a relevantních klíčových slov a souvislých popisů pro středověké rukopisy na základě vstupních strukturovaných popisů a obrazových dat.
- Zlepšení třídění a vyhledávání v katalogu Manuscriptoria.
- Nové pokročilé funkce implementované do uživatelského interface díky využití nově získaných metadat a popisů.
- Zvýšení efektivity práce s databází pro odborníky i veřejnost.
- Disseminace poznatků po dokončení výzkumného záměru (*workshop* pro paměťové instituce).
-

Způsob řešení

Záměr bude řešen v těsné součinnosti smluvních stran.

Fáze 1: Analýza a návrh

1. Navrhnout obecné *workflow* pro automatizované generování klíčových slov a popisů pro agregovaný obsah, včetně návrhu datových struktur a návrh popis úprav prostředí Manuscriptoria
2. Analyzovat vhodnost využití umělé inteligence pro generování klíčových slov a popisů na vybraném vzorku dat a metadat, tzn. pro obrazy i metadata, kdy přednostně s ohledem na kapacitu a ceny budeme v případě obrazů pracovat s obrazy stránek/folií které jsou odkazovány z popisných metadat
3. V případě neuspokojivých výsledků prověřit alternativní dostupné modely nebo AI
4. V případě uspokojivých výsledků vyladit prompty k dosažení co nejlepších výsledků.
5. V rámci plnění dojde k formulování use cases pro potřeby posouzení úspěšné realizace záměrů a akceptace výsledků ze strany objednatele.

Fáze 2: Integrace AI nástrojů a aktualizace *workflow* zpracování agregovaného obsahu

6. Vyvinout integrační body pro výměnu informací s AI
7. Úpravy *workflow* v systému Manuscriptoria
8. Provedení strukturálních změn v datovém modelu Manuscriptoria

Fáze 3: Úpravy UI

9. Návrh nových dílčích nástrojů a funkcionalit v systému Manuscriptoria
10. Kustomizovat SW Gulliver podle návrhu

Fáze 4: Testování, poloprovoz a validace

- Provést důkladné testování a ladění systému
- Validovat obohacující metadata za pomoci odborníků na středověké rukopisy.
- Uvedení systému do režimu poloprovozu

Fáze 5: Nasazení, workshop

- Nasadit finální verzi nástrojů a *workflow* v rámci provozu Manuscriptoria 4
- Provést *workshop* a sdílet zkušenosti se zainteresovanými institucemi (v anglickém jazyce)

Orientační rozpis jednotlivých fází projektu

Fáze	MD Poskytovatel
1: Analýza a návrh	44
2: Integrace AI nástrojů a aktualizace <i>workflow</i> zpracování agregovaného obsahu	48
3: Úpravy UI	32
4: Testování, poloprovoz a validace	10
5: Ostrý provoz, workshop	4