beyond the concept

Návrh riešenia vyhľadávania – ÚPVS

Verzia: 1.0
Dátum: 26. 9. 2025
Vypracoval: Tomáš Majerov – CTO

Dataconcept s.r.o,
Prievozská 4D, 821 09 Bratislava | ilike@dataconcept.sk | www.dataconcept.digital

1. Úvod

Tento dokument popisuje návrh implementácie full-textového vyhľadávania pre ÚPVS. Riešenie kombinuje CMS Drupal a vyhľadávací engine Meilisearch. Cieľom je zlepšiť presnosť, relevantnosť a používateľskú skúsenosť pri vyhľadávaní. Dokument zároveň poskytuje opis technických riešení jednotlivých požiadaviek.

2. Navrhované riešenie požiadaviek

2.1. Oprava preklepov

Táto funkcionalita je už implementovaná prostredníctvom mechanizmu typo tolerance v Meilisearch. Tento mechanizmus využíva výpočet tzv. Levenshteinovej vzdialenosti medzi hľadaným výrazom a slovami v indexe a na základe toho povoľuje menšie odchýlky.

Nastavené sú dynamické hranice podľa dĺžky slova: pri krátkych slovách (do 4 znakov) je povolený jeden preklep, pri dlhších slovách (5 a viac znakov) maximálne dva. Nad osem znakov zostáva hranica na úrovni dvoch preklepov. Tento prístup zabezpečuje, že krátke slová nebudú generovať príliš veľa nerelevantných výsledkov.

Tolerancia preklepov sa neuplatňuje pri prefixovom vyhľadávaní (autocomplete), pretože by viedla k neúmernému množstvu návrhov. Rovnako sa nepoužíva pri vyhľadávaní na základe filtrov alebo pri číselných hodnotách (napr. IČO).

Pri hodnotení výsledkov majú vždy prednosť tie, ktoré sa zhodujú presne. Výsledky s preklepmi sa zobrazia až za nimi, pričom o poradí rozhoduje interné ranking pravidlo „typo“.

Príklady správania:

- dopyt *„dôchdca“* nájde výsledok *„dôchodca“* (1 preklep),
- dopyt *„eletrnicka schranka“* nájde výsledok *„elektronická schránka“* (viacero preklepov, stále v tolerancii),
- dopyt *„zmlv“* nájde výsledok *„zmluva“* (1 preklep).

Funkcionalita dopĺňania návrhov pri písaní (autocomplete s prefix search) je rovnako už implementovaná a funguje od zadania minimálne troch znakov.

2.2. Synonymický slovník

V Drupale sa vytvorí nový obsahový typ „**Synonymum**“, ktorý bude slúžiť na správu kľúčových výrazov a ich alternatívnych pomenovaní. Redaktori doň budú zapisovať kanonický pojem (hlavný tvar slova alebo frázy) a k nemu priraďovať zoznam synoným, ktoré majú viesť na rovnaký výsledok vo vyhľadávaní. Takto bude možné zabezpečiť, že používateľ nájde relevantný obsah aj vtedy, keď použije iný, ale významovo podobný výraz. Dáta zo synonymického slovníka sa budú pravidelne exportovať do **Meilisearch** pomocou naplánovaného **cron jobu**, ktorý zabezpečí aktuálnosť a konzistenciu indexu.

2.3. Automatické dopĺňanie a návrhy pri písaní

Táto požiadavka je už v systéme **implementovaná** a **plne funkčná**. Automatické dopĺňanie a návrhy pri písaní sa spúšťajú po zadaní minimálne troch znakov do vyhľadávacieho poľa. Mechanizmus je postavený na prefix search v Meilisearch, čo znamená, že systém vyhľadáva všetky slová začínajúce na zadaný reťazec. Používateľ tak okamžite vidí relevantné možnosti, napríklad po zadaní „elek…“ sa zobrazia návrhy ako „elektronické služby“, „elektronická schránka“ či „elektronické podanie“. Návrhy sú generované na základe obsahu indexu a uprednostňujú sa tie, ktoré sú frekventovanejšie alebo boli často vyhľadávané. Vďaka tomu sa používateľovi výrazne skracuje čas potrebný na nájdenie správneho výsledku a zvyšuje sa celkový komfort práce so systémom.

2.4. Rozšírené vyhľadávanie, pokročilé hľadanie a filtre

Požiadavka je plne realizovateľná v rámci možností **Meilisearch** prostredníctvom mechanizmu **filterableAttributes**. Pri indexovaní obsahu z Drupalu sa budú do Meilisearch prenášať doplnkové polia predstavujúce metadáta dokumentu. Medzi tieto atribúty patrí typ obsahu (článok, elektronická služba, životná situácia), dátum poslednej aktualizácie, ako aj informácie o inštitúcii alebo úrade a o územnej pôsobnosti.

Atribúty *inštitúcia/úrad a územná pôsobnosť* budú dopĺňané v Drupale len pre vybrané typy obsahu (node typy), ktoré budú definované dohodou. Pre správne fungovanie bude potrebné pripraviť **číselníky úradov a územných pôsobností**, ktoré zabezpečia jednotnosť a konzistenciu údajov. Obsah týchto číselníkov a spôsob ich správy bude stanovený na základe dohody.

Používateľské rozhranie poskytne jednoduché možnosti na aplikovanie týchto filtrov prostredníctvom výberových polí, prepínačov alebo dátumových komponentov.

V praxi to znamená, že ak používateľ zadá kľúčové slovo „schránka“, môže si zároveň nastaviť, aby sa zobrazili len elektronické služby, prípadne len služby konkrétneho úradu (vybraného zo zoznamu číselníka) alebo obmedzené podľa časového obdobia. Takýmto spôsobom sa výrazne zvýši presnosť a relevantnosť vyhľadávania, pričom systém zostane rýchly a používateľsky komfortný.

2.5. Sémantické vyhľadávanie a pochopenie kontextu + Podpora morfologickej normalizácie a rodových variantov

Na pokrytie tejto požiadavky bude zavedené sémantické vyhľadávanie, ktoré umožní prácu s významom slov a fráz namiesto čistej textovej zhody. V praxi to znamená, že pri dotaze „postup na zmenu zdravotnej poisťovne“ systém správne identifikuje a ponúkne výsledky súvisiace s elektronickým podaním pre zmenu zdravotnej poisťovne, aj keď sa tieto slová doslovne nezhodujú s názvom služby. Zároveň sa bude brať do úvahy aj kontext predošlých dotazov, takže vyhľadávač dokáže prepojiť súvisiace dopyty a zvýšiť tým relevanciu ponúknutých výsledkov.

Plánované riešenie počíta s využitím NLP mikroslužby pre spracovanie dotazov a ich transformáciu na kľúčové slová, synonymá a morfologické varianty. Takto obohatený dotaz sa následne odosiela do Meilisearch, ktorý zabezpečí rýchle a presné vyhľadávanie.

Ako alternatívu je možné uvažovať o využití komerčnej služby Meilisearch Cloud Pro, ktorá poskytuje plne spravované prostredie s prioritnou podporou, analytikou a monitoringom. Cena služby začína od 300 USD mesačne a zahŕňa 250 000 vyhľadávaní a 1 milión dokumentov.

2.5.1. NLP microservice – účel a funkcionalita

NLP microservice bude samostatný komponent systému, ktorou úlohou je **spracovať dotaz používateľa ešte predtým, ako sa odošle do Meilisearch**. Cieľom je zvýšiť relevanciu výsledkov tým, že sa voľne formulovaný text rozloží na **kľúčové slová**, rozšíri o **synonymá** a doplní o **morfologické a rodové varianty**.

Hlavné funkcie

1. Predspracovanie dotazu

- odstránenie nadbytočných znakov,
- normalizácia (malé písmená, diakritika),
- stoplist (vylúčenie bezvýznamových slov ako „na“, „do“, „je“).

2. Morfologická analýza a lematizácia (SK)

- identifikácia lemma (základného tvaru) pre každé slovo,
- získanie morfologických rysov (rod, číslo, pád) pre potreby rozšírenia.

3. Neutralizácia rodu/čísla/pádu

- pre každý term sa vytvorí **neutrálny dotazový reprezentant** (lemma),
- doplnia sa **ekvivalenty rodu** (napr. *dôchodca* ↔ *dôchodkyňa*, *študent* ↔ *študentka*) a vybrané flexné tvary, aby pokryli bežné formulácie,
- pravidlá sa opierajú o:
  - **morfologické paradigmy** (suffixové páry typu *-ca/-kyňa*, *-ent/-entka*),
  - **výnimky/lexikón** (kurátne spravovaný zoznam ťažko odvoditeľných párov v Drupale).

4. Extrakcia kľúčových slov / fráz

- výber nosných pojmov z dlhších viet (KeyBERT/YAKE alebo TF-IDF),
- rozpoznanie viacslovných pojmov (napr. *zdravotná poisťovňa*).

5. Rozšírenie o synonymá (Drupal slovník)

- doplnenie kanonických/alternatívnych výrazov (redakčne spravované),
- kombinácia so **rodovými ekvivalentmi** (synonymá sa rovnako neutralizujú a párujú).

6. Vytvorenie dotazu pre Meilisearch

- boolean/phrased query: *(lemma OR rodové ekvivalenty OR synonymá)* pre každý kľúčový pojem,
- rešpektovanie filtrov (typ, úrad/inštitúcia, pôsobnosť, dátum).

Technológie a implementácia

- **Jazykové knižnice**:
    - **Stanza** (Stanford NLP) alebo **UDPipe 2** – morfologická analýza a lematizácia pre slovenčinu.
    - **KeyBERT** alebo **YAKE** – extrakcia kľúčových slov.
- **Integrácia so synonymickým slovníkom**: REST API napojené na Drupal.
- **Framework**:
    - **FastAPI (Python)** na implementáciu mikroslužby.
    - Bude vystavený endpoint, ktorý prijme dotaz a vráti obohatený variant.
- **GPU nie je nutná** – plánované je použitie **lightweight modelov optimalizovaných pre CPU**.

Príklad spracovania dotazu

Vstup: "postup na zmenu zdravotnej poisťovne"

Spracovanie:

- normalizácia → "postup zmena zdravotná poisťovňa"
- lematizácia → "postup zmena zdravotná poisťovňa"
- extrakcia kľúčových slov → "postup", "zmena", "zdravotná poisťovňa"

- rozšírenie synonymami → "postup | návod | spôsob", "zmena | prehlásenie", "zdravotná poisťovňa | zdravotná inštitúcia"

Výstup do Meilisearch:
("postup" OR "návod" OR "spôsob") AND ("zmena" OR "prehlásenie") AND ("zdravotná poisťovňa" OR "zdravotná inštitúcia")

2.6. Indexácia súborov a metaúdajov

Požiadavka bude riešená prostredníctvom samostatnej **mikroslužby pre spracovanie príloh**, ktorá zabezpečí extrakciu textu z nahraných dokumentov a jeho následné odoslanie do indexu Meilisearch. Predpokladá sa, že väčšina dodávaných súborov bude textovo extrahovateľná, a preto **nebude potrebné OCR spracovanie**.

Proces bude fungovať tak, že pri nahratí alebo aktualizácii súboru v Drupale sa vygeneruje úloha pre mikroslužbu, ktorá pomocou nástroja **Apache Tika** extrahuje textový obsah. Spolu s tým sa do indexu uložia aj dôležité **metadáta** (názov súboru, typ, dátum nahratia, väzba na nadradený obsah). Výsledkom je, že používateľ dokáže vo vyhľadávaní nájsť nielen články a samotný webový obsah, ale aj informácie priamo v prílohách.

V Meilisearch budú extrahované dáta zaradené do štruktúrovaných polí, pričom text dokumentov sa použije ako **searchableAttribute** a základné metadáta ako **filterableAttributes**. Z hľadiska relevantnosti bude obsah príloh hodnotený nižšie ako názvy alebo súhrny článkov, aby sa zamedzilo tomu, že rozsiahle dokumenty potlačia primárny obsah.

Používateľské rozhranie zároveň poskytne jasnú informáciu o tom, že zhoda bola nájdená v prílohe – napríklad formou označenia *„Zhodu v prílohe: názov_súboru.pdf“*.

V prípade, že by sa v budúcnosti vyskytli súbory bez textovej vrstvy (napr. naskenované dokumenty), riešenie umožní doplnenie OCR spracovania ako voliteľného rozšírenia, avšak v aktuálnom návrhu sa s touto funkcionalitou **nepočíta**.

2.7. Testovanie / akceptačné kritériá

Testovanie bude rozdelené do viacerých častí, pričom pre každú z nich budú pripravené konkrétne scenáre, metriky a spôsob realizácie.

- **Testy s preklepmi a synonymami** – bude pripravených minimálne **20 testovacích scenárov** (10 preklepových, 6 synonymických, 4 kombinované). Testovanie bude vykonané manuálne podľa vopred pripravených dotazov a paralelne aj automatizovane ako integračné testy, ktoré overia konzistentnosť výsledkov. Akceptačným kritériom je, aby systém vždy vrátil relevantné výsledky a správne uprednostnil bezchybné zhody.

- **Testy filtrovania a rozšíreného vyhľadávania** – bude pripravených minimálne **24 scenárov** kombinujúcich typ obsahu, inštitúciu/úrad, územnú pôsobnosť a dátum aktualizácie. Tieto testy budú realizované formou **automatizovaných UI testov** (napr. pomocou Cypress alebo Playwright) a manuálne overované testerom. Kritériom je, aby boli výsledky konzistentné s nastavenými filtrami a facety zobrazovali korektné počty.

- **Testy výkonu** – budú realizované napríklad pomocou nástrojov **k6** a **Locust**. Scenáre budú zahŕňať bežnú záťaž (50 RPS), navýšenú záťaž (100 RPS), extrémne špičky (200 RPS), ako aj dlhodobé testy stability (2–4 hodiny). Testovať sa bude na korpusoch o veľkosti 300 tisíc, 600 tisíc a 1 milión dokumentov. Metriky:

    - latencia,
    - chybovosť.

  Testy budú prebiehať na **staging prostredí**, ktoré zrkadlí produkčné nastavenia.

Testy použiteľnosti (UX) – bude pripravených minimálne **6 používateľských úloh** (vyhľadanie e-služby, filtrovanie podľa úradu, použitie dátumu, vyhľadanie s preklepom atď.), ktoré otestuje **8 účastníkov**. Testovanie bude prebiehať formou **moderovaných používateľských testov**, kde sa bude sledovať, či používateľ dokáže bez pomoci úspešne

splniť úlohy a či rozumie výsledkom a filtrom. Akceptačným kritériom je úspešnosť aspoň 80 % účastníkov a spokojnosť s jasnosťou filtrov minimálne 4/5.

2.8. Odhad náročnosti

Sekcia	Názov / Funkcionalita	Popis prác	Odhad MD
2.1	Oprava preklepov	Implementované v Meilisearch, , testovanie a doladenie	**3**
2.2	Synonymický slovník	Typ obsahu v Drupale, formulár, validácia, export do Meilisearch (cron), testovanie	**6**
2.3	Automatické dopĺňanie a návrhy pri písaní	Implementované, testovanie a doladenie	**4**
2.4	Rozšírené vyhľadávanie a filtre	Rozšírenie indexu o metadáta, číselníky v Drupale, UI komponenty, testovanie	**12**
2.5	Sémantické vyhľadávanie a morfologická normalizácia	NLP mikroslužba: FastAPI, Stanza/UDPipe, KeyBERT/YAKE, Redis cache, testovanie	**24**
2.6	Indexácia súborov a metaúdajov	Mikroslužba (Apache Tika), integrácia s Drupalom, odosielanie metadát, UI pre výsledky, testovanie	**12**
2.7	Testovanie / akceptačné kritériá	Scenáre, automatizované UI testy (Cypress/Playwright), záťažové testy (k6/Locust), UX testy (6 úloh × 8 účastníkov)	**27**