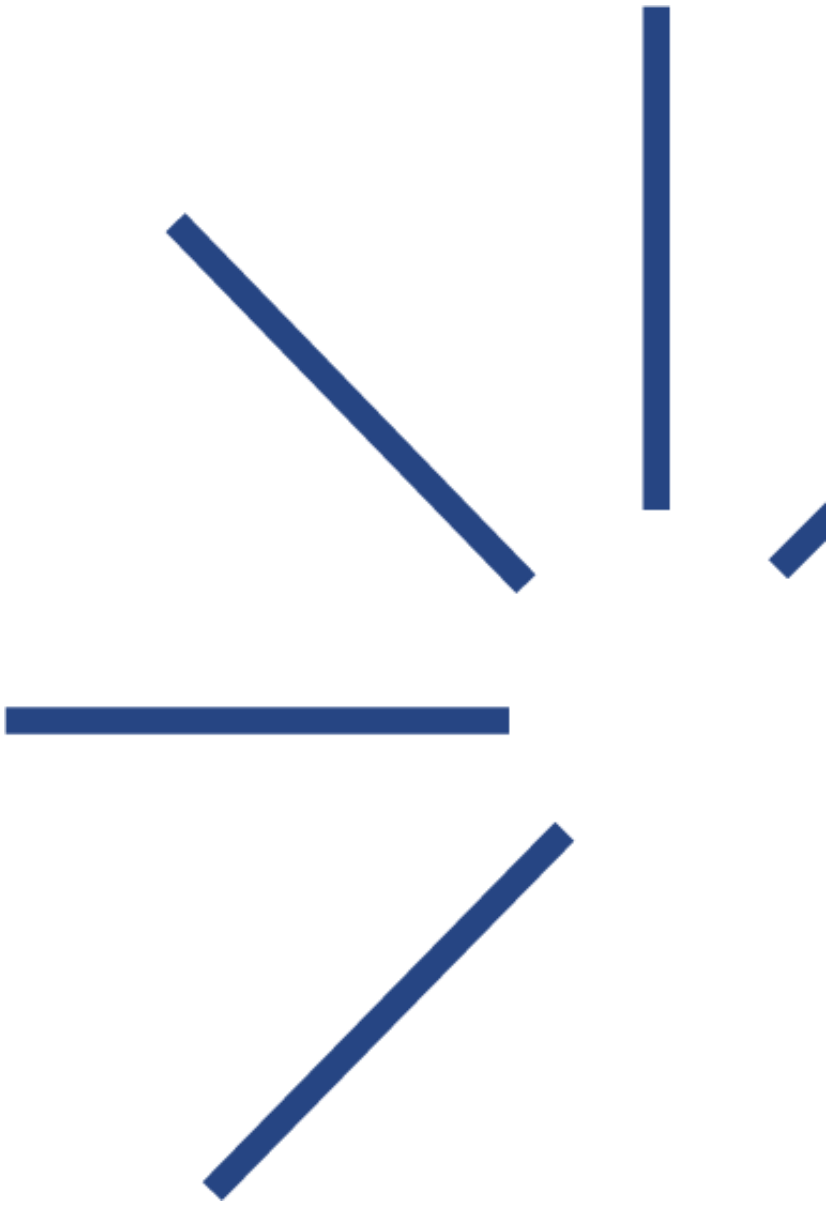
D15 – Návrh repozitára



Projekt financovaný Európskou úniou Next GenerationEU prostredníctvom Plánu obnovy a odolnosti SR v rámci projektu č. 09I05-03-V02-00049.

ÚVOD

Projekt „**Automatizácia analýzy právnych textov na základe strojového učenia**“ (ďalej len „**ALTAML**“) predstavuje kľúčovú iniciatívu zameranú na integráciu inovatívnych prístupov v oblasti spracovania údajov a následnej analýzy, so zameraní na oblasť práva, ktorá zahŕňa aj právo informačných a komunikačných technológií. Cieľom tohto projektu je vyvinúť a overiť účinné metódy automatizovanej analýzy právnych textov pomocou techník strojového učenia. To zahŕňa najmä vývoj nástrojov, ktoré uľahčujú spracovanie a analýzu veľkých objemov údajov vo forme rôznych právnych dokumentov, ako aj extrakciu relevantných informácií (atribútov) z takýchto dokumentov vrátane identifikácie kľúčových pojmov, odkazov na právne predpisy a ďalších atribútov.

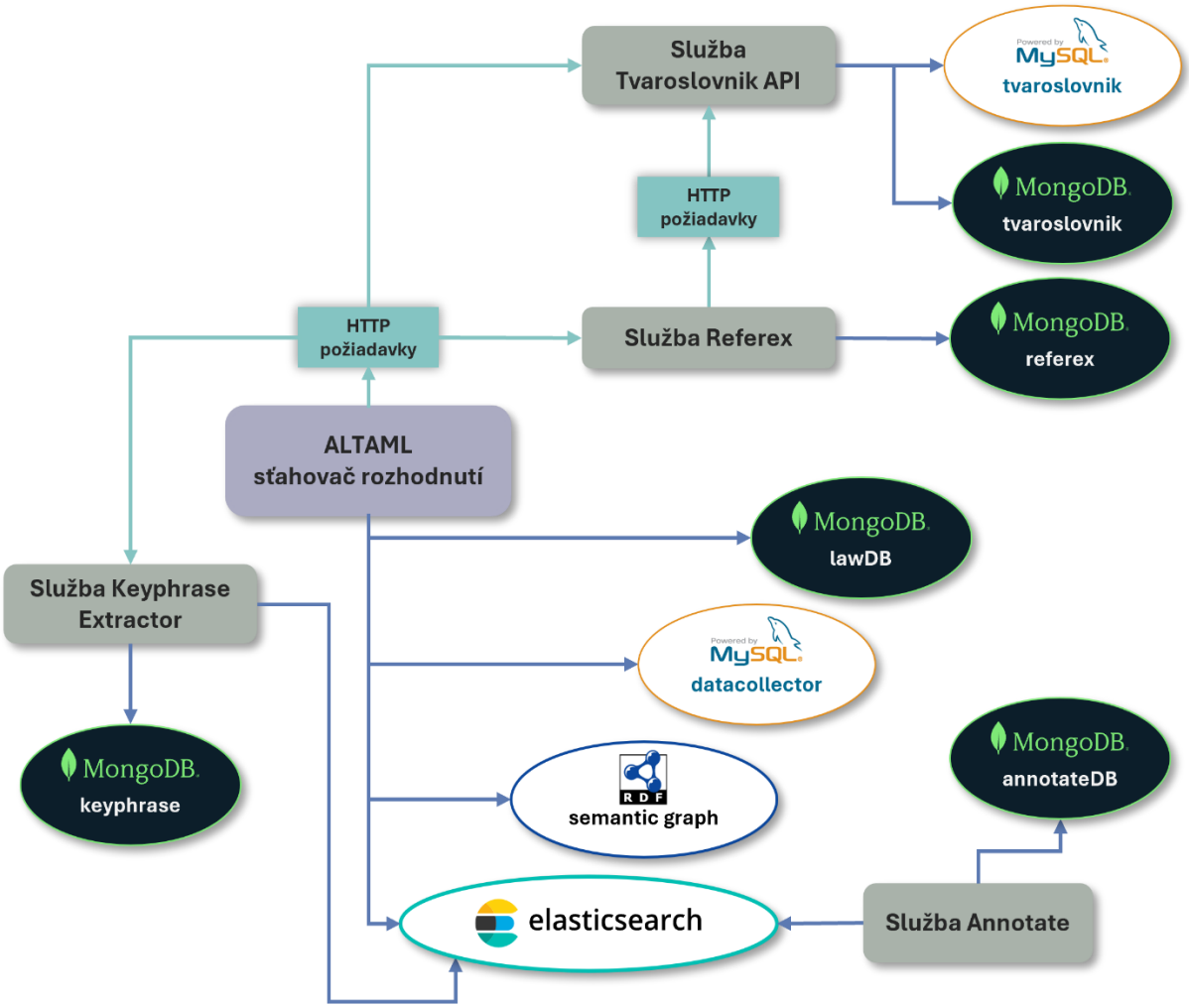
Cieľom projektu ALTAML je teda prispieť k efektívnejšiemu prístupu k právnym informáciám a k zrýchleniu právnych procesov, čím sa zabezpečí vyššia miera právnej istoty a zlepší sa dostupnosť právnych textov, výsledkov ich analýzy a relevantných právnych informácií pre odbornú aj širokú verejnosť.

V rámci pracovného balíka KPB3, bol vytvorený **D15 – Návrh repozitára**. Tento dokument sa zameriava na návrh dátovej architektúry systému pre automatizované spracovanie, obohacovanie a sprístupňovanie súdnych rozhodnutí. Opisuje navrhované komponenty systému a ich dátové repozitáre vrátane mechanizmov pre riadenie spracovateľskej pipeline, verziovanie spracovania a ukladanie obohatených dát.

Návrh kombinuje relačné, dokumentové a vyhľadávacie databázy tak, aby podporoval jednotlivé fázy spracovania právnych textov. Osobitná pozornosť je venovaná návrhu repozitárov pre jazykové spracovanie textov, extrakciu právnych odkazov a identifikáciu kľúčových právnych pojmov. Cieľom navrhovanej architektúry je vytvoriť škálovateľný a rozšíriteľný základ pre ďalší rozvoj systému.

1. Návrh systému a príslušných repozitárov

Navrhovaný systém predstavuje modulárnu architektúru určenú na automatizované spracovanie, obohacovanie a sprístupňovanie súdnych rozhodnutí. Celkový pohľad na návrh architektúry a repozitárov je znázornený na Obrázok 1.



Obrázok 1 - Návrh repozitára

Základom systému je komponent ALTAML sťahovač rozhodnutí, ktorý má zabezpečovať získavanie rozhodnutí z externých zdrojov, ich spracovanie, verziovanie a ukladanie do vhodných dátových repozitárov.

Na jazykové spracovanie textov má byť využívaná služba Tvaroslovník API, poskytujúca lematizáciu a gramatické informácie pre slovenský jazyk.

Extrakciu štruktúrovaných odkazov na právne predpisy a súdne rozhodnutia má realizovať samostatná služba Referex, zatiaľ čo identifikáciu kľúčových právnych pojmov má zabezpečovať služba Keyphrase Extractor s využitím sémantických

vektorových reprezentácií. Finálne vyhľadávacie a analytické rozhranie systému má tvoriť služba Annotate, umožňujúca vyhľadávanie nad rozhodnutiami a ich anotáciu v používateľskom nástroji. Navrhovaná architektúra je postavená na oddelených komponentoch a repozitároch, čo má umožniť škálovateľnosť, opakovateľnosť spracovania a postupné rozširovanie systému.

2. Komponent ALTAML sťahovač rozhodnutí

ALTAML sťahovač rozhodnutí je komponent zodpovedný za automatizované sťahovanie, spracovanie a obohacovanie súdnych rozhodnutí z externých zdrojov, najmä z API Ministerstva spravodlivosti (Slov-lex). Zabezpečuje kontrolu verzií rozhodnutí, konverziu dokumentov do textovej podoby a postupné aplikovanie nezávislých extrakčných metód. Výsledkom jeho činnosti sú štruktúrované a obohatené rozhodnutia pripravené na ukladanie.

ALTAML sťahovač rozhodnutí využíva viacero repozitárov podľa fázy spracovania.
MySQL slúži ako riadiaci repozitár na evidenciu rozhodnutí, ich verzií a technických metadát. **MongoDB** sa používa na ukladanie priebežných a finálnych obohatených verzií rozhodnutí vrátane histórie extrakčných metód. **Elasticsearch** predstavuje finálny vyhľadávací repozitár určený na fulltextové, atribútové a semantické vyhľadávanie nad súdnymi rozhodnutiami.
Okrem toho ALTAML sťahovač rozhodnutí ukladá vybrané extrahované vzťahy aj do **sémantického grafu**, ktorý neobsahuje celé texty rozhodnutí, ale iba štruktúrované entity a ich vzájomné väzby.

MongoDB databáza *lawDB*

Databáza lawDB slúži ako hlavné úložisko finálne obohatených rozhodnutí v internom dátovom modeli. Rozhodnutie sa do databázy uloží až vtedy, keď je kompletne spracované – teda po konverzii PDF na text a po vykonaní všetkých extrakčných krokov (napr. právne predpisy, odkazy na iné rozhodnutia, kľúčové pojmy a ďalšie atribúty). Finálne obohatené rozhodnutia sa ukladajú do kolekcie **decision_clean_raw**.

V tejto databáze sa zároveň uchováva informácia o verziách použitých metód, takže pre jedno súdne rozhodnutie môže existovať viacero verzií finálneho objektu podľa toho, aké verzie extrakčných metód boli použité. Tento prístup umožňuje návrat k starším verziám v prípade, že sa nová extrakčná metóda ukáže ako chybná. Zároveň umožňuje spustiť novú alebo upravenú extrakčnú metódu iba nad už finálne

spracovanými rozhodnutiami a uložiť výsledok pod novou verziou, bez potreby opätovného spúšťania celého spracovateľského procesu.

Po vykonaní všetkých extrakčných a spracovateľských krokov a uložení výsledkov do databázy sa vyberie jedna konkrétna verzia rozhodnutí, ktorá sa považuje za aktuálnu. Súdne rozhodnutia s touto verziou sú následne použité pre ďalšie systémy, najmä pre vyhľadávanie s pomocou databázy Elasticsearch a služby Annotate.

MySQL databáza *datacollector*

MySQL **datacollector** slúži ako riadiaca a referenčná databáza ALTAML sťahovača rozhodnutí, ktorej hlavnou úlohou je koordinovať spracovanie súdnych rozhodnutí. Neuchováva samotný obsah rozhodnutí, ale údaje potrebné na rozhodovanie o tom, čo má byť spracované, v akej verzii a v akom stave sa dané rozhodnutie nachádza.

Kľúčovým prvkom je tabuľka **raw_decision**, ktorá predstavuje centrálnu evidenciu všetkých rozhodnutí identifikovaných v externých zdrojoch. Pre každé rozhodnutie uchováva interný identifikátor, informácie o zdroji dát, dátum poslednej aktualizácie v zdrojovom systéme, ako aj technické údaje o fyzickom uložení dokumentu (napr. cesta k PDF súboru).
Vďaka týmto informáciám je možné rozlíšiť nové rozhodnutia od aktualizovaných, zabrániť redundantnému spracovaniu a presne sledovať stav spracovania.

MySQL zároveň plní úlohu referenčného úložiska zdieľaných a dlhodobo platných údajov, ktoré sa počas spracovania opakovane využívajú. Ide najmä o tabuľku **sud**, obsahujúcu normalizované informácie o súdoch (typ súdu, kraj, okres), a tabuľku **data_source_schema**, v ktorej sú uložené historické aj aktuálne JSON schémy response z externých API.
Tieto informácie sa ukladajú zámerne, aby nebolo potrebné pri každom spracovaní rozhodnutia opakovane volať ďalšie endpointy externého systému, čím sa znižuje zaťaženie serverov Ministerstva spravodlivosti a zároveň sa skracuje čas spracovania.

Okrem toho sa v MySQL ukladajú aj číselníkové údaje, ktoré sa využívajú v iných častiach projektu, hlavne na filtrovanie súdnych rozhodnutí. Ide o tabuľky **povaha_rozhodnutia, oblast_pravnej_upravy, podoblast_pravnej_upravy** a **forma_rozhodnutia**, ktoré obsahujú hodnoty zodpovedajúce názvu tabuľky.

Fulltextové a vektorové vyhľadávacie úložisko Elasticsearch

Elasticsearch slúži ako finálna vyhľadávacia vrstva systému nad súdnymi rozhodnutiami. Do databázy sa ukladajú rozhodnutia vybranej verzie z MongoDB kolekcie decision_clean_raw, pričom každé súdne rozhodnutie je v Elasticsearch reprezentované jedným objektom.

Táto vrstva nie je určená na uchovávanie histórie ani verzií rozhodnutí, ale výhradne na ich efektívne vyhľadávanie a prístup. Elasticsearch je využívaný na fulltextové vyhľadávanie v textoch rozhodnutí, filtrovanie podľa štruktúrovaných a extrahovaných atribútov (napr. súd, spisová značka, ECLI, dátum rozhodnutia) a na pokročilé vyhľadávacie scenáre.

Pri indexovaní sa používa vlastný analyzér so slovenským Hunspell slovníkom, ktorý zohľadňuje rôzne morfologické tvary slov, čo je kľúčové pri práci so slovenskými právnymi textami.

3. Služba Tvaroslovník API

Tvaroslovník API je REST služba na spracovanie slovenčiny – používa sa na lematizáciu (prevod slov do základného tvaru) a/alebo na vrátenie gramatických kategórií pre zadané slovo, či množinu slov. Na tieto úlohy využíva dve databázy: **MongoDB** a **MySQL**.

MySQL databáza *tvaroslovnik*

MySQL obsahuje jedinú tabuľku **lemma**, ktorá tvorí jadro lematizácie. V stĺpci *tvar* je uložený nelematizovaný tvar slova zo vstupu, v stĺpci *lemma* jeho základný tvar a stĺpec *pocet* reprezentuje frekvenciu výskytu daného základného tvaru v slovenčine podľa jazykových štatistík. Ak pre jeden tvar existuje viacero možných lém, uprednostní sa tá, ktorej základný tvar je v slovenčine používanejší (má vyšší *pocet*).

Z dôvodu výkonu pri spracovaní dlhších textov je tabuľka optimalizovaná pomocou primárneho kľúča, ktorý umožňuje rýchle vyhľadávanie podľa tvaru.

MongoDB databáza *tvaroslovnik*

MongoDB v rámci Tvaroslovník API slúži na uchovávanie dát, ktoré nie sú priamo súčasťou lematizačnej tabuľky. Kolekcia **kategorie** obsahuje gramatické kategórie slov, pričom identifikátor dokumentu (dokument – jeden objekt v rámci kolekcie) predstavuje konkrétny tvar slova a uložené údaje sú využívané endpointom vracajúcim gramatické informácie.

Kolekcia **unlemmatized_words** eviduje slová, ktoré sa nepodarilo lematizovať, pretože sa nenachádzajú v MySQL tabuľke **lemma**. Tieto slová sú následne dopĺňané jazykovedcami, aby bolo možné ich lematizovať pri ďalšom spracovaní.

4. Služba Referex

Referex je samostatná služba určená na automatickú extrakciu odkazov z právnych textov, najmä na slovenské právne predpisy, predchádzajúce súdne rozhodnutia, rozhodnutia Európskeho súdu pre ľudské práva (ESĽP), Súdneho dvora Európskej únie (SDEÚ) a Úradu na ochranu osobných údajov (ÚOOÚ).

Táto služba využíva výhradne databázu MongoDB ako svoj dátový repozitár. Nasledujúce MongoDB kolekcie sú v systéme Referex využívané na podporu presnej identifikácie právnych predpisov a súdnych rozhodnutí v právnych textoch.

MongoDB databáza *referex*

Kolekcia **zakony_aliases** obsahuje úplné názvy zákonov spolu s ich identifikátormi a slúži na vyhľadanie zhody názvu zákona priamo v texte právneho predpisu. Na tento účel nadväzuje kolekcia **dict_lemmatized_laws**, ktorá obsahuje tie isté názvy zákonov v lematizovanej podobe, aby bolo možné hľadať zhody aj v lematizovanom texte rozhodnutí.

Keďže v súdnych rozhodnutiach sa často nepoužívajú celé názvy zákonov, ale iba ich skratky, využíva sa kolekcia **law_aliases**. V nej je identifikátorom skratka názvu zákona a atribút *law_ids* obsahuje zoznam identifikátorov zákonov, ktoré môžu byť touto skratkou označené. Táto kolekcia slúži na zúženie množiny kandidátov pri identifikácii zákona v prípade, že je v texte použitá iba skratka.
Kolekcia **law_aliases_lemmatized** je obsahovo rovnaká, avšak identifikátory sú lematizované, keďže môžu obsahovať aj celé slová. Pri extrakcii sa využívajú obe kolekcie a výsledok sa vyberá podľa najbližšej zhody s textom právneho predpisu.

Kolekcia **law_validities** obsahuje dokumenty, v ktorých identifikátor je číslo zákona (napr. 300/2005 pre Trestný zákon) a atribút *validities* uchováva jednotlivé časové intervaly platnosti jeho verzií. Pri extrakcii právnych predpisov sa používa na priradenie správnej verzie zákona k extrahovanému odkazu.

Kolekcia **codebook_courts** sa používa pri extrakcii odkazov na iné slovenské súdne rozhodnutia. Obsahuje názvy súdov a ich typy, ktoré pochádzajú z číselníka súdov

získaného z ALTAML sťahovača rozhodnutí, a umožňuje spoľahlivú identifikáciu súdu uvedeného v texte rozhodnutia.

5. Služba Keyphrase Extractor

Keyphrase Extractor je samostatná služba určená na automatickú extrakciu kľúčových právnych pojmov zo súdnych rozhodnutí. Pri výpočte kľúčových pojmov využíva samotný text rozhodnutia, ako aj extrahované odkazy na právne predpisy a iné súdne rozhodnutia spolu s ich textami.

Na prácu s kandidátnymi pojmami a na sémantické porovnávanie vektorových reprezentácií využíva databázy **MongoDB** a **Elasticsearch**.

MongoDB databáza *keyphrase-extractor*

MongoDB slúži ako referenčné úložisko kandidátnych právnych pojmov. V kolekcii *candidate_terms* je uložená množina kandidátnych pojmov rozdelených na obsahové a procesné, z ktorých sa vyberajú kľúčové pojmy pre konkrétne rozhodnutie.

Vektorové úložisko Elasticsearch

Elasticsearch je využívaný ako vektorová databáza obsahujúca vektorové reprezentácie textov právnych predpisov, súdnych rozhodnutí a ich vzájomných vzťahov. Na základe ich vektorového porovnania sa vyhodnocuje sémantická blízkosť kandidátnych pojmov k spracovávanému rozhodnutiu, čo umožňuje výber najrelevantnejších kľúčových právnych pojmov.

6. Služba Annotate

Annotate je REST služba, ktorá sprístupňuje endpointy na vyhľadávanie súdnych rozhodnutí a zároveň poskytuje kompletnú podporu pre anotáciu dát v anotačnom nástroji.

Vyhľadávanie je realizované prostredníctvom Elasticsearch dopytov, ktoré Annotate service preberá priamo z frontendového anotáčneho nástroja a vykonáva ich nad indexom súdnych rozhodnutí.

Fulltextové a vektorové vyhľadávacie úložisko Elasticsearch

Pôvodne sa uvažovalo o návrhu vlastného dopytovacieho jazyka a middleware vrstvy pre komplexné filtrovanie právnych dokumentov, avšak tento prístup bol nahradený využitím natívneho dopytovacieho mechanizmu Elasticsearch. Ten sa ukázal ako dostatočný a zároveň umožňuje kombinovať klasické atribútové a fulltextové vyhľadávanie so sémantickým vyhľadávaním založeným na vektorových reprezentáciách uložených priamo v Elasticsearch.

MongoDB databáza *annotate*

MongoDB databáza je využívaná na vytváranie dátových sád súdnych rozhodnutí, spravovanie anotačných úloh, priraďovanie úloh anotátorom a ukladanie samotných anotácií.