D2 – Metóda spracovania právnych textov

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 zameraním 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 KPB1 bol vytvorený výstup **D2 - Metóda spracovania právnych textov** – určenie metód spracovania právnych textov pre akýkoľvek subjekt, ktorý má záujem o spracovanie právnych textov.

Obsah

Úvod **1**

1 Spracovanie súdnych rozhodnutí **3**

2 Hromadne sťahované rozhodnutia **5**

3 Manuálne sťahované rozhodnutia **9**

4 XSLT transformácia **10**

5 Nástroj na export metadát a tvorbu dátových balíkov **11**

6 Záverečné zhodnotenie **12**

1 Spracovanie súdnych rozhodnutí

Spracovanie súdnych rozhodnutí v rámci projektu ALTAML pozostáva zo štyroch vrstiev:

(a) Získavanie dát (crawl/API/manuál),

(b) Transformácia (PDF→XML, XSLT, prelinkovanie),

(c) Kvalita a obohatenie (validácie, mapovanie metadát, inferencia vzťahov),

(d) Publikácia a export (CMS, dátové balíky, HTML výstupy).

Každá vrstva má definované vstupy, výstupy, kontroly a SLA reakcie na chyby. Kľúčové princípy: jednotná štruktúra XML, reprodukovateľnosť (skriptované pipeline), trasovateľnosť (logy a audit), a interoperabilita (ECLI, štandardy metadát).

V rámci projektu ALTAML získala spoločnosť Wolters Kluwer (ďalej len "WK") rozhodnutia prostredníctvom dvoch metód: automatizovaného sťahovania prostredníctvom zväčša programových robotov a manuálneho stiahnutia.

Počas procesu **hromadného získavania súdnych rozhodnutí** boli okrem samotných PDF dokumentov získané aj podrobné informácie o jednotlivých rozhodnutiach, ktoré slúžili ako metadáta a dopĺňali informácie v rozhodnutiach.

PDF dokumenty boli najprv hromadne konvertované do formátu XML skriptom vyvinutým spoločnosťou WK, ktorý využíval nástroje poppler-utils a XSL transformácie na technicky precízne prevedenie dát.

V nadväznom kroku aplikácia skonštruovaná spoločnosťou WK pridala odkazy na relevantné právne predpisy, súdne rozhodnutia, ako aj legislatívu Európskej únie, čím zabezpečila úplnosť a referenčnú hodnotu týchto dokumentov pre ďalšie použitie.

Dokumenty vo formáte XML boli následne importované do systému Content Managment System (CMS, kde prebehli formálne procesy validácie za účelom zabezpečenia správnosti a integrity údajov podľa stanovených štandardov a požiadaviek.

Získané detaily jednotlivých rozhodnutí boli potom použité na systematické generovanie metadát súvisiacich s týmito rozhodnutiami. Tento proces zahŕňal namapovanie a porovnanie získaných údajov voči existujúcim kategóriám a terminologickým slovníkom, čím bola zabezpečená konzistentnosť a správnosť informácií.

Okrem toho, ďalšie metadáta boli vytvárané v následnom procese z už vytvorených XML súborov rozhodnutí, umožňujúc detailné prehľad o každom rozhodnutí prostredníctvom komplexného a systematizovaného uchovávania informácií.

Pri **manuálnom procese získavania rozhodnutí** dochádzalo k stiahnutiu samostatných dokumentov rozhodnutí, alebo kompletných súborov, klasifikovaných ako zošity rozhodnutí, pričom sa využíval formát PDF pre zachovanie integrity a ochrany obsahu. Tento postup si vyžaduje presnú kontrolu a verifikáciu jednotlivých stiahnutých dokumentov, aby sa zabezpečila ich správnosť a úplnosť v kontexte rozhodovacej dokumentácie.

V rámci uvedeného procesu boli PDF dokumenty prevedené do formátu XML prostredníctvom interného skriptu vytvoreného spoločnosťou WK. Následne, pomocou špecifickej aplikácie taktiež vyvinutej WK, boli do textu implementované hypertextové odkazy na súvisiace právne predpisy, súdne rozhodnutia a legislatívu Európskej únie. Pri dokumentoch, ktoré boli dostupné vo formáte obrazových PDF, bol použitý OCR nástroj, konkrétne Abbyy FineReader, na digitalizáciu a konverziu obsahu do textovej podoby.

Transformované rozhodnutia vo formáte XML boli importované do systému správy obsahu (CMS), v ktorom prebehli nevyhnutné procesy validácie, ktorými sa zabezpečila ich technická a obsahová správnosť.

Metadáta manuálne stiahnutých súdnych rozhodnutí boli individuálne zaznamenané do systému riadenia obsahu (CMS) a získané z PDF dokumentu alebo z webovej stránky pôvodného zdroja.

2 Hromadne sťahované rozhodnutia

Harmonogram ťahov je nastavený na pravidelné dávky (cron). Pri každom ťahu sa vykoná kontrola zmenených alebo nových GUID/ID a deduplikuje sa podľa hash súboru a spisovej značky.

Validácie po importe do CMS zahŕňajú: (1) schému XML, (2) prítomnosť povinných metadát (druh, dátum, spisová značka, ECLI ak je dostupné), (3) konzistenciu prelinkovania na právne predpisy a judikatúru, (4) anti-duplicačné pravidlá.

2.1 A.1. Rozhodnutia zo zdroja justice.gov.sk

Ide o rozhodnutia všeobecných súdov, ktoré sú uverejnené na oficiálnej webovej stránke Ministerstva spravodlivosti SR.

URL:

- https://www.justice.gov.sk/sudy-a-rozhodnutia/sudy/rozhodnutia (vyhľadávací formulár)
- https://obcan.justice.sk/pilot/api/ress-isu-service/swagger-ui/index.html#/rozhodnutie (API)

Podrobná informácia o rozhodnutí vrátane jeho metadát je uložená vo formáte JSON a je prístupná na stiahnutie prostredníctvom API endpointu dostupného na adrese

https://obcan.justice.sk/pilot/api/ress-isu-service/v1/rozhodnutie/{guid}.

Potom sa stiahnu textové súbory súdnych rozhodnutí vo formáte PDF prostredníctvom adries URL, ktoré sú špecifikované v poli details[„document“][„url“].

Príklad URL adresy s podrobnosťami o rozhodnutí (JSON):

https://obcan.justice.sk/pilot/api/ress-isu-service/v1/rozhodnutie/6e12997b-519c-4799-99fd-a3dbcf8db7cb:47305804-6213-45b8-ae44-0d009d23e52a

Príklad URL s textom PDF:

https://obcan.justice.sk/content/public/item/47305804-6213-45b8-ae44-0d009d23e52a

Príklad detailu rozhodnutia zo zdroja:



2.2 A.2. Rozhodnutia zo zdroja nsud.sk

Ide o zdroj rozhodnutí Najvyššieho súdu SR ktoré sú uverejnené na oficiálnej webovej stránke tohto súdu.

URL: https://www.nsud.sk/rozhodnutia/

Metadáta sú získavané z HTML stránky obsahujúcej podrobné informácie o konkrétnom právnom rozhodnutí, ako aj hypertextový odkaz na pripojený súbor vo formáte PDF obsahujúci samotné rozhodnutie.

Medzi metadáta extrahované priamo z HTML patrí názov rozhodnutia, dátum jeho vydania a referenčné číslo prípadu. Súčasne sa určité metadáta extrahujú prostredníctvom vytvoreného XML dokumentu, ktoré zahŕňajú konkrétnu kategorizáciu rozhodnutia, identifikačné číslo súboru, identifikátor európskej judikatúry a relevantné vzťahy k existujúcim právnym aktom a iným judikatúrnym textom.

2.3 A.3. Rozhodnutia zo zdroja nssud.sk

Na webovej stránke Najvyššieho správneho súdu SR sú rozhodnutia rozdelené do piatich hlavných kategórií:

- **rozhodovacie činnosti v správnych veciach:**

https://www.nssud.sk/rozhodovacia-cinnost-vo-veciach-spravnych/, https://www.nssud.sk/wp-admin/admin-ajax.php?action=get_wdtable&table_id=70 (údaje JSON)

- **rozhodovacie činnosti v disciplinárnych veciach**

https://www.nssud.sk/rozhodovacia-cinnost-vo-veciach-disciplinarnych/, https://www.nssud.sk/wp-admin/admin-ajax.php?action=get_wdtable&table_id=67 (údaje JSON)

- **rozhodovacie činnosti vo volebných veciach**

https://www.nssud.sk/rozhodovacia-cinnost-vo-veciach-volebnych/, https://www.nssud.sk/wp-admin/admin-ajax.php?action=get_wdtable&table_id=61 (údaje JSON)

- **rozhodnutia kompetenčného senátu**

https://www.nssud.sk/rozhodnutia-kompetencneho-senatu/, https://www.nssud.sk/wp-admin/admin-ajax.php?action=get_wdtable&table_id=64 (údaje JSON)

- **rozhodnutia veľkého senátu**

https://www.nssud.sk/rozhodovacia-cinnost-velkeho-senatu/

V rámci procesu získavania metadát z elektronických dokumentov sú extrahované nasledujúce informácie z JSON formátu: názov dokumentu, spisová značka, register a tematická oblasť. Na druhej strane, z XML formátu sú extrahované metadáta, ktoré zahŕňajú typ dokumentu, identifikačné číslo spisu, európsky identifikátor judikatúry (ECLI) a dátum vydania rozhodnutia.

2.4 Rozhodnutia zo zdroja ustavnysud.sk

Ide o rozhodnutia Ústavného súdu, ktoré sú uverejnené na oficiálnej webovej stránke tohto súdu.

URL: https://www.ustavnysud.sk/rozhodovacia_cinnost/rozhodnutia

V priebehu počiatočného štádia je realizovaná extrakcia metadát z objektov dokumentov, ktoré boli stiahnuté. V následnom kroku sa uskutočňuje extrakcia metadát z textov rozhodnutí spracovaných vo formáte XML. Konkrétne, metadáta extrahované z dokumentových objektov zahŕňajú typ rozhodnutia, dátum vydania rozhodnutia, európsky identifikátor judikatúry (ECLI) a spisovú značku. Je však preferované, aby bola spisová značka overená prostredníctvom formátu XML. Z textov vo formáte XML sa získavajú metadáta, medzi ktoré patrí spisová značka, informácie týkajúce sa vzťahu rozhodnutia k príslušným právnym predpisom a existujúcej judikatúre.

Príklad dokumentového objektu zo zdroja:


```
{
    "documents": [
        {
            "docType": "USSR_DECISION",
            "content": null,
            "title": "Rozhodnutia - Uznesenia o spojení vecí",
            "index": 0,
            "documentId": "132d0d4b-9c0e-4dec-9319-876a02128d23",
            "mkDocumentType": "Rozhodnutia - Uznesenia o spojení vecí",
            "mkDateOfDecision": "04/09/2025 00:00:00",
            "mkRSAPNumberOfFile": "PLs. ÚS 25/2025",
            "mkTypeOfDecision": [
                "Uznesenie o spojení"
            ],
            "mkYearOfLawReports": null,
            "contentType": null,
            "mkDecisionInTermsOf": [
                "§ 62 zákona o ústavnom súde v spojení s § 166 ods. 1 CSP"
            ],
            "mkLawReportsNumber": null,
            "mkParentIdDecision": null,
            "extension": null,
            "size": null,
            "mkVolumeOfLawReports": null,
            "mkAffectedLegalRegulation": null,
            "mkClauseTitle": null,
            "mkClauseText": null,
            "mkTimePeriodZNaU": null,
            "mkWebTitle": null,
            "mkECLI": "ECLI:SK:USSR:2025:PLs.US.25.2025.1",
            "mkDateOfLegalForce": "04/09/2025 00:00:00",
            "mkPublicationDate": null,
            "mkTypeOfProposer": null,
            "mkTypeOfProceeding": "procesné konanie ÚS SR",
            "mkFormOfDecision": "Uznesenie",
            "mkTypeOfNegotiation": [
                "Prerokovanie vo veci"
            ],
            "mkResultOfNegotiation": [
                "spája na spoločné konanie",
                "nadobúda právoplatnosť a vykonateľnosť"
            ],
            "mkCause": null,
            "mkWordRegister": [
                "spojenie vecí"
            ],
            "mkMaterialRegister": [
                "konanie pred ústavným súdom-použitie ustanovení Civilného sporového poriadku"
            ],
            "mkComplainedLegalRegulation": null,
            "mkJudgeReporter": "uboš Szigeti",
            "mkDifferentView": null,
            "mkUnderage": null,
            "mkRVPNumberOfFile": "928/2025",
            "mkFileReference": null,
            "mkReferences": null,
            "mkIncludeToZnaU": false,
            "mkClarificationOfLegalRegulation": null,
            "mkEntryDate": null,
            "mkFormOfEntry": null,
            "mkTypeOfEntry": null,
            "mkDecisionInTermsOfForSort": "§ 62 zákona o ústavnom súde v spojení s § 166 ods. 1 CSP"
        }
```

3 Manuálne sťahované rozhodnutia

Proces získavania tohto typu obsahu predstavuje zložitejšiu úlohu, ktorá je charakteristická prevažnou potrebou manuálneho zásahu pri vykonávaní väčšiny súvisiacich činností, ktoré nie sú plne automatizované.

Manuálny workflow: (1) identifikácia zošita/rozhodnutia, (2) sťahovanie PDF, (3) OCR v ABBYY FineReader pri obrazových PDF, (4) transformácia PDF→XML, (5) doplnenie metadát a kontrola kvality, (6) import do CMS. Pri zošitoch sa jednotlivé rozhodnutia segmentujú na základe titulkov, číslovania a vzorov.

V rámci našich iniciatív predstavujeme komplexný prehľad URL adresných zdrojov zameraných na tieto konkrétne typy obsahu:

3.1 Zbierka rozhodnutí a stanovísk Najvyššieho súdu SR

Tento dokument sústreďuje vybrané rozhodnutia vydané Najvyššieho súdu SR, ktoré sú redakčne upravené a usporiadané. Publikácia je koncipovaná formou zošita, čo znamená, že obsahuje viaceré rozhodnutia v jednom súbore.

URL: https://www.nsud.sk/zbierka-stanovisk-a-rozhodnuti/

3.2 Zbierka rozhodnutí a stanovísk Najvyššieho správneho súdu SR

Ide o vybrané rozhodnutia Najvyššieho správneho súdu SR, ktoré sú redakčne upravené a usporiadané. Publikácia je koncipovaná formou zošita, čo znamená, že obsahuje viaceré rozhodnutia v jednom súbore.

URL: https://www.nssud.sk/zbierka-stanovisk-a-rozhodnuti-najvyssieho-spravneho-sudu-slovenskej-republiky/

3.3 Zbierka rozhodnutí Ústavného súdu SR

Ide o vybrané rozhodnutia Ústavného súdu SR, ktoré sú redakčne upravené a usporiadané. Sú vydávané buď jednotlivé, alebo raz ročne súhrnne v jednom PDF dokumente.

URL: https://www.ustavnysud.sk/sk/rozhodovacia_cinnost/zbierka_rozhodnuti

4 XSLT transformácia

Za účelom realizácie projektu ALTAML spoločnosť Wolters Kluwer vyvinula skript, ktorý implementuje XSLT transformačný proces na konverziu XML dokumentov do formátu HTML.

Formát HTML je optimalizovaný na prehliadanie v internetových prehliadačoch a zabezpečuje vyššiu zrozumiteľnosť pre ďalších používateľov.

XSLT šablóny definujú hlavičku, štruktúru rozhodnutí. Štýly sú navrhnuté tak, aby boli čitateľné v prehliadači a zároveň konzistentné s interným dizajn manuálom.

5 Nástroj na export metadát a tvorbu dátových balíkov

V rámci implementácie projektu bola spoločnosťou Wolters Kluwer vyvinutá softvérová aplikácia umožňujúca export metadát z výsledkov SQL dotazov predložených užívateľom, pričom aplikácia slúži aj na generovanie dátových balíkov.

Užívateľ má možnosť formulovať súbor otázok prostredníctvom SQL dotazov, ktoré sa vzťahujú na jednotlivé témy, respektíve domény projektu ALTAML.

Výsledkové množiny, vygenerované na základe uvedených dotazov, dodržiavajú nasledovnú štruktúru metadát: Id; aspi_id; autor; spzn; druh; datum; pramen; pub_cislo; full_pramen; oblast; relation; rejsrik; rok; porc; vydani.

Táto štruktúra metadát umožňuje tvorbu zostáv, ktoré sú plne v súlade s požiadavkami zadávateľov projektu a ktoré vedú ku konkrétnym prehľadom dátových balíkov, vrátane samotných dátových balíkov obsahujúcich html dokumenty.

Výstupné metadáta (Id; aspi_id; autor; spzn; druh; datum; pramen; pub_cislo; full_pramen; oblast; relation; rejsrik; rok; porc; vydani) umožňujú skladbu prehľadov a integráciu do downstream analytiky. Vznikajú HTML balíky s prelinkovanými dokumentmi.

6 Záverečné zhodnotenie

Spracovanie súdnych rozhodnutí predstavuje komplexný proces, ktorý prepája automatizované technológie, štruktúrované dátové modely a odborné doménové znalosti.

V rámci projektu **ALTAML** sme vytvorili jednotný, metodicky podložený a technicky konzistentný postup, ktorý umožňuje zabezpečiť vysokú kvalitu, spoľahlivosť a opakovateľnosť celého workflow – od prvotného získavania dokumentov až po tvorbu dátových balíkov pripravených na ďalšie analytické a aplikačné využitie.

Kľúčovými prvkami úspechu sú:

- dôsledné uplatňovanie validačných mechanizmov,
- štandardizácia formátov a metadát,
- transparentné prelinkovanie na legislatívne a judikatúrne zdroje,
- riadenie rizík a kontinuálne zlepšovanie procesov.

Zavedené postupy dávajú projektu ALTAML potenciál poskytovať kvalitné, dôveryhodné a technicky kompatibilné justičné dáta. Tento dokument zároveň slúži ako ucelený metodický rámec, podľa ktorého je možné proces ďalej rozvíjať, optimalizovať a prispôsobovať novým požiadavkám.