

# Zpracování datové sady evidovaných knihoven

Tento dokument popisuje návrh postupu pro automatizované zpracování, uložení a popis datové sady knihoven evidovaných Ministerstvem kultury ČR.

## 1. Způsob automatického stahování dat

Data jsou na stránce Ministerstva kultury k dispozici jako soubor ve formátu Microsoft Excel (.xlsx). Pro zajištění automatického stahování bych zvolila následující postup:

Vytvoření skriptu v programovacím jazyce Python: Python je pro tento účel ideální díky široké podpoře knihoven pro práci s webem a daty. Pro samotné stažení souboru z dané URL adresy by klíčovou knihovnou byla Requests. Skript pro stažení dat je uložen v samostatném souboru skript\_stazeni\_dat.py v tomto repozitáři.

Proces skriptu:

- Skript pomocí knihovny Requests odešle HTTP GET požadavek na přímou URL adresu XLSX souboru.
- Zkontroluje stavový kód odpovědi. Pokud je kód 200 (OK), obsah odpovědi uloží na definované místo, Ukládání s datem stažení zajistí jednoduchou historizaci a sledovatelnost.

Při psaní skriptu je vhodné zvážit situace, které při stahování dat mohou nastat, a ošetřit je. Skript lze rozšířit například tak, aby:

- podporoval různé typy souborů a formátů (CSV, JSON, XML, ZIP, případně i PDF)
- zahrnoval rozšířenou detekci chyb a výjimek
- poslal upozornění e-mailem na případné chyby nebo selhání při stažení
- nabízel možnost automatického rozbalení a předzpracování dat
- umožňoval verzování dat a porovnání změn

Skript lze rozšířit o další části do funkcí/modulů. Cílem je, aby byl skript více univerzální a vhodný pro opakované použití.

- download\_file()
- validate\_response()
- save\_file()
- send\_notification()
- log\_error()

Automatizace: Pro pravidelné spouštění skriptu (např. jednou měsíčně, aby byla data aktuální) bych využila nástroj operačního systému jako Plánovač úloh ve Windows nebo Cron na Linuxu/macOS. Existují samozřejmě i sofistikovanější nástroje pro orchestraci dat.

## **2. Transformace a úpravy datové sady pro zvýšení interoperability**

Surová data v Excelu jsou zřídka kdy připravena pro přímé použití. Pro zvýšení jejich hodnoty a možnosti napojení na další zdroje (interoperabilita) bych provedla následující kroky, opět s využitím Pythonu a knihovny Pandas:

Načtení dat: Načtení dat z xlsx souboru do datového rámce (DataFrame) v Pandas.

Čištění a standardizace (Data Cleaning):

- Názvy sloupců: Přejmenování sloupců na strojově čitelné názvy bez diakritiky a mezer (např. Název knihovny -> nazev\_knihovny).
- Datové typy: Kontrola a správné nastavení datových typů pro jednotlivé sloupce (např. IČO jako text, aby se neodstranily úvodní nuly; počet svazků jako číslo).
- Textová data: Sjednocení textových hodnot – odstranění nadbytečných mezer, sjednocení velikosti písmen (např. Lowercase)
- Chybějící hodnoty: Analýza chybějících hodnot (NULL/NaN) a rozhodnutí o strategii jejich ošetření (ponechání, smazání řádku nebo doplnění, pokud je to možné).

Obohacení a transformace pro interoperabilitu:

Geokódování: Adresa je v datech uložena jako text. Pro strojové zpracování a vizualizace (např. na mapě) je klíčové převést adresu na geografické souřadnice.

Využití identifikátorů: Datová sada obsahuje IČO. Tento unikátní identifikátor je klíčový pro napojení na další české registry, jako je Administrativní registr ekonomických subjektů (ARES). Pomocí IČO je možné automaticky doplnit další údaje – např. přesný název subjektu, právní formu, oficiální sídlo atd. Tím se data obohatí a zároveň validují.

Strukturování adresy: Sloupec s adresou rozdělit na jednotlivé části (ulice, číslo popisné/orientační, obec, PSČ) do samostatných sloupců. To usnadní filtrování a spojování s jinými datovými sadami, které mají adresu také takto strukturovanou.

Výstupní formát:

Po transformaci bych data uložila do standardizovaného, otevřeného formátu, jako je CSV (pro maximální kompatibilitu). Excel (.xlsx) není pro další strojové zpracování vhodný.

## **3. Zajištění trvalého a bezpečného uložení dat**

Ideálním místem pro uložení transformovaných a vyčištěných dat je relační databáze.

Struktura: V databázi bych vytvořila tabulku s jasně definovaným schématem, které odpovídá vyčištěným datům (názvy sloupců, datové typy, omezení jako NOT NULL).

Integrita dat: Databáze zajistí datovou integritu – např. unikátnost IČO, dodržování datových typů.

Dotazování: Data v databázi jsou snadno dostupná pro další analýzy, vizualizace nebo napojení na jiné aplikace pomocí jazyka SQL.

Bezpečnost:

Řízení přístupu: Vytvoření dedikovaného databázového uživatele s omezenými právy (např. pouze pro čtení dat pro analytiku, práva pro zápis pouze pro automatizovaný skript). Hesla a přístupové údaje by nikdy nebyly uloženy přímo ve skriptu, ale ve specializovaném a zabezpečeném úložišti.

Zálohování: Nastavení pravidelných a automatických záloh celé databáze. Zálohy by měly být testovány, zda jsou obnovitelné.

Trvalost a verzování:

Při každém novém stažení a zpracování dat bych starší verzi dat v databázi buď aktualizovala (pokud chceme mít jen aktuální stav) nebo přidala nové řádky s časovým razítkem stažení. To by umožnilo sledovat vývoj datové sady v čase.

Zbývá se rozhodnout, zda archivovat původní stažené soubory (.xlsx) pro případnou budoucí potřebu zpětné kontroly.

#### **4. Zajištění popisu datové sady pomocí Czech Core Metadata Model (CCMM)**

Aby byla datová sada srozumitelná a dohledatelná pro ostatní uživatele, je nezbytné vytvořit kvalitní metadata. Czech Core Metadata Model je založen na mezinárodních standardech (DCAT, Dublin Core) a je pro tento účel ideální.

Vytvořila bych metadatový záznam (např. ve formátu JSON-LD), který by popisoval námi vytvořenou, vyčištěnou a obohacenou datovou sadu. Záznam by obsahoval například tyto položky:

dct:title: "Vyčištěná a geokódovaná datová sada evidovaných knihoven v ČR"

dct:description: "Datová sada vznikla automatizovaným zpracováním, čištěním a geokódováním dat z Evidence knihoven, kterou spravuje Ministerstvo kultury ČR. Obsahuje informace o knihovnách, jejich zřizovatelích, adresách a typech, obohacené o přesné geografické souřadnice a propojení na registr ARES."

dct:publisher: "Ekonomicko-správní fakulta, Masarykova univerzita" (protože fakulta je vydavatelem této nové, odvozené datové sady).

dct:source: Uvedení odkazu na původní zdroj dat na webu Ministerstva kultury.

dcat:keyword: "knihovny", "kultura", "geodata", "otevřená data", "registr"

dct:modified: Datum poslední aktualizace dat (vkládáno automaticky skriptem).

dct:spatial: "Česká republika" (jako text) a zároveň vymezení pomocí souřadnic (bounding box) získaných z geokódovaných dat.

dcat:distribution: Popis konkrétních způsobů, jak se k datům dostat. Například:

Jedna distribuce ve formátu CSV (s odkazem na stažení souboru).

Druhá distribuce jako API (pokud by data byla zpřístupněna přes databázový endpoint).

U každé distribuce by byl uveden formát (text/csv), licence (např. CC BY 4.0) a odkaz na stažení.

Tento metadatový záznam bych uložila jako soubor (metadata.jsonld) vedle samotných dat a zveřejnila v rámci GitHub repozitáře.