**Příloha č. 4 - Technická specifikace**

**DL4DH – vývoj softwarového řešení**

Cílem projektu DL4DH je vývoj nového softwarového řešení, které umožní extenzivní využití a vytěžování dat z digitálních knihoven provozovaných v systému Kramerius (<https://system-kramerius.cz/>) pro potřeby digitálního humanitního výzkumu. Software bude vyvíjen jako open source pod licencí GNU GPL v3. Pro vývoj a správu zdrojového kódu bude využita služba GitHub.

Základem vyvíjeného řešení je software **DL4DH Feeder** provozovaný jako webová služba určená pro přímé využití vědeckými pracovníky. Pro ukládání metadat a obohacených textových dat, které nelze získat ze  systému Kramerius, bude vytvořen pomocný systém **Kramerius plus**. Součástí vyvíjeného řešení bude i možnost propojení webové služby DL4DH Feeder a systému Kramerius plus s externími databázemi a nástroji a jejich využití při obohacování metadat a textových dat nebo při vizualizaci dat vytěžených ze systému Kramerius. Nový nástroj **TEI Converter** umožní export dat ze systému Kramerius a databáze Kramerius plus ve formátu TEI (viz <https://tei-c.org/guidelines/p5/>).

Schéma řešení:



**Modul Kramerius plus (Kramerius+)**

Rozšiřující modul Kramerius+ bude sloužit pro uchování obohacených textových dat a metadat, které jsou uloženy v systému Kramerius, jedná se na např. o lemmatizaci a automaticky rozpoznané entity v plných textech, záznamy z knihovních systémů, odkazy na autority a bibliografické záznamy, apod.). Tyto údaje se využijí pro kvalitnější vyhledávání digitálních dokumentů, k filtraci jejich obsahu a umožní obsah Krameria vědecky zpracovávat, případně exportovat pro další použití.

Kramerius+ by měl fungovat jako externí modul systému Kramerius, měl by mít vlastní API a měl by komunikovat s jádrem Krameria i s DL4DH Feederem. Kramerius+ bude využívat existující plné texty publikací, které jsou uložené v systémech Kramerius a které obvykle vznikly pomocí OCR (ve formátu prostého textu nebo ALTO, viz <https://www.loc.gov/standards/alto/>).

Specifikace funkcí modulu Kramerius+

* Jednotlivé záznamy v Krameriu+ budou svázány s dílčími objekty v Krameriu na úrovni stránky publikace. Pokud to výhledově půjde, lze uvažovat o navázání přímo na dílčí objekty v ALTO (např. odstavce), které však není v současnosti dostupné pro všechny dokumenty v Krameriu. Řešení bude v tomto ohledu univerzální.
* V Krameriu+ budou uložena metadata rozšiřující dosavadní popis dokumentů v systému Kramerius o informace podstatné pro pokročilé dotazy a obsahové vyhledávání v dokumentech. Zvolená technologie musí umožnit efektivní prohledávání těchto obohacených dat (řádově miliony metadatových položek a desítky milionů stran textu).
* Kramerius+ se bude starat o obohacování metadat a textových dat voláním interních, popř. externích služeb s požadovanými parametry. Součástí popisu služeb bude i jejich účel (např. lemmatizace, vyhledání záznamů v národních autoritách ap.). Zajistí také uložení metadat o volané službě (název služby, verze, nastavené parametry, datum volání a vrácení výsledku).
* Výstupy z externích služeb spolu s aktuální podobou textových dat ve formátu XML TEI předá Kramerius+ nástroji **TEI Converter**, který zajistí konverzi a integraci těchto výstupů do formátu TEI. Výsledný dokument se uloží v Krameriu+.
* Validní TEI export pro API komunikující s DL4DH Feederem bude umět složit Kramerius+ s využitím nástroje TEI Converter.
* Při zpracování výstupů pro DL4DH Feeder se v maximální míře budou využívat metadata ze systému Kramerius. Existující údaje se tak nebudou zbytečně duplikovat v pomocné databázi, pokud pro to nebudou existovat jiné závažné důvody (zejména efektivita a rychlost zpracování dotazů).
* Obohacené záznamy se budou za účelem vývoje, testování a poloprovozu vytvářet pro cíleně zvolené skupiny dat z Krameria. Úzké hrdlo při velkých objemech je zejména dostupný prostor na úložištích. Pro vlastní provoz tak bude důležitá možnost vytvářet fronty uživatelských požadavků pro zpracování publikací do obohaceného formátu. Při dostatečných kapacitách může být obohaceno vše v Krameriovi.
* Technologie pro pomocnou databázi Kramerius+ bude zvolena v průběhu implementační analýzy.
* Data Krameria+ by mělo být možné zpřístupnit jako Linked Open Data.
* Kramerius+ musí podporovat správu přístupů v souladu s nastavením licenčních podmínek v základním Krameriu.
* Bude probíhat automatická aktualizace metadat na základě zdrojového záznamu z knihovního systému a báze národních autorit (přebírání MARC záznamů).
* Měla by probíhat automatická aktualizace metadat/dat mezi Krameriem a Krameriem+
* Součástí řešení bude interní analytika využívání systému kvůli vylepšování funkcí i textových dat a metadat.
* K obohacení metadat budou využita metadata z archivních balíčků a zároveň bude možné se přes perzistentní identifikátory provázat na zdrojový obrázek v plném rozlišení

Specifikace obohacování dat

Cílem procesu obohacování je doplnit metadata a plný text (vzniklý OCR) ze systému Kramerius o strukturní značky, jazykovou analýzu, rozpoznané entity a další údaje. Obohacování bude zpravidla plně automatizovaný proces, který využije zejména externí služby a nástroje, schopné zpracovat data z Krameria v požadované formě a rozsahu. V případech, kdy je to možné, mohou být využity i externí výpočetní kapacity (např. infrastruktura LINDAT/CLARIAH-CZ).

Vlastní nástroje pro zpracování dat budou vyvinuty pouze v případech, které jsou kritické pro chod systému a nemají existující funkční alternativu.

Integrace systému s externími i interními nástroji musí být modulární: proces zpracování bude možné administrátorskými zásahy upravovat a např. připojovat nové nástroje a upravovat parametry obohacování (nastavení vstupního a výstupního formátu apod.).

Součástí systému budou také údaje o průběhu obohacení textových dat a metadat: označení externí, popř. interní aplikace, webová/lokální adresa, verze aplikace, nastavení parametrů pro zpracování.

**Mezi základní požadavky pro obohacení patří:**

Integrace funkcí pro jazykové zpracování *(využití volně dostupné infrastruktury LINDAT/CLARIAH-CZ):*

* Tokenizace (rozdělení na slova a věty), lemmatizace (určení základního tvaru slova) a určení slovního druhu (např. nástrojem UDPipe)
	+ pro texty v češtině, němčině, latině, příp. dalších jazycích.
* Rozpoznání a určení typu pojmenovaných entit (např. pomocí nástroje NameTag).
	+ Přinejmenším pro češtinu, další jazyky zatím nejisté.
* Převod do TEI - bude zajištěn nově vyvinutým nástrojem TEI Converter, jehož specifikace je součástí této zadávací dokumentace.

Cílem mimo jiné je automaticky analyzovat metadata již dostupná v ALTO (nechceme tato metadata nově vytvářet, pouze efektivně využít). Objem takto zpracovatelných dat poroste tak, jak se bude rozšiřovat obsah digitální knihoven. Zpracování jednotlivých funkcí bude závislé na vstupní analýze a nemusí být všechny realizovány, pokud se k tomu nepodaří nalézt technické prostředky a jejich vývoj nebude v rámci projektu proveditelný. Jde např. o tyto funkce:

* Identifikace obrázků na stránkách, jejich polohy a popisků.
* Rozlišení veršovaného a prozaického textu.
* Rozpoznání struktury textu (sloupce atd.) a podílu prázdného místa na stránce (v %).
* Rozpoznání a správná interpretace tabulek, např. pomocí nástroje Tabula (<https://tabula.technology/>).
* Analýza typografie.

**Další možnosti pro obohacení**

Cílem je pokusit se najít služby, které by dokázaly naplnit následující požadavky, případně vyhodnotit, nakolik je možné v rámci projektu jejich úpravou dosáhnout požadovaného cíle a poskytnou takový výsledek pro Feeder. Bude implementováno podle časových možností a podrobnější vstupní analýzy v průběhu řešení projektu. Výsledkem může být i test s vyhodnocením nasaditelnosti v budoucnu.

Příkladem jsou tyto funkce:

* Klasifikace obrázků podle základního typu obsahu - fotografie / mapa / kresba (atp.). Lze vycházet též z návěstí (popisků) obrázků.
* Identifikace rukopisných přípisků.
* Rozpoznávání bibliografických odkazů, resp. obecně vylepšení práce s referencemi v rámci seznamu literatury, např. pomocí nástroje Grobid (<https://github.com/kermitt2/grobid>).
* Segmentace slovníkových a encyklopedických textů, např. pomocí nástroje Grobid-Dictionaries (<https://github.com/MedKhem/grobid-dictionaries>).
* Obohacování o článková metadata za využití oborové nebo článkové bibliografie (např. Česká literární bibliografie, Bibliografie dějin Českých zemí, Bibliografie české lingvistiky), vč. prolinkování na konkrétní článek a zpět.

### Specifikace funkcionality a rozhraní nástroje „TEI Converter“ pro obohacování dat z digitálních knihoven provozovaných v systému Kramerius

TEI Converter je softwarové řešení, které komunikuje s modulem Kramerius+, aby rozpoznaný text digitalizátu obohatil o relevantní data a metadata a aby uživateli poskytl sadu dat v požadované kvalitě (metadata, úroveň obohacení textu) ve formátu XML TEI. Jeho funkce budou využívat správci modulu Kramerius+ a prostřednictvím nástroje DL4DH Feeder, jej budou využívat vědečtí pracovníci v oblasti digitálních humanitních věd. TEI Converter nebude mít vlastní grafické uživatelské rozhraní.

TEI Converter musí poskytovat kvalitně vytvořené a zdokumentované API (s využitím Swaggerhubu, Apiary nebo podobných nástrojů), které umožní modifikaci jednotlivých kroků v procesu obohacování a poskytování dat.

Popis požadované funkčnosti

1. Při vytváření nebo aktualizaci záznamů v systému Kramerius+ vytvoří TEI Converter výchozí podobu dokumentu TEI (tj. elementu <text>) na základě výstupů OCR (ve formátu prostého textu nebo ALTO).
2. Převod výstupů z volaných služeb na odpovídající prvky a atributy TEI, případně sloučení s dokumentem TEI z předchozích kroků.
3. Sestavení kompletního dokumentu, resp. sady dokumentů ve formátu TEI na základě metadat ze systému Kramerius a dat, popř. metadat z modulu Kramerius+. TEI Converter zajistí např. vytvoření hlavičky (<teiHeader>) s metadaty o digitalizátu a jeho zpracování a obohacování nebo spojení samostatných stran publikace do jednoho dokumentu (v elementu <tex>). Dokumenty budou obsahovat metadata, text a jeho obohacení v takové úrovni, kterou si nastavil uživatel prostřednictvím nástroje DL4DH Feeder.

**Specifikace funkcionality a rozhraní nástroje “DL4DH Feeder” pro vytěžování dat z digitálních knihoven provozovaných v systému Kramerius**

DL4DH Feeder je softwarové řešení napojené na systém Kramerius a na modul Kramerius+, které zajišťuje selekci, vizualizaci a následný export dat do některého z požadovaných formátů (CSV, TSV, XML, JSON, ALTO, PAGE), umožňující další (zejména) strojové zpracování. Software by měly mít možnost nasadit knihovny využívající systém Kramerius (viz registrkrameriu.mzk.cz/). Uživatelé, kteří budou využívat jeho funkce, budou vědečtí pracovníci v oblasti digitálních humanitních věd. DL4DH Feeder bude mít vlastní uživatelské rozhraní umožňující plné využití jeho funkcí.

DL4DH Feeder musí poskytovat kvalitně vytvořené a zdokumentované API (s využitím Swaggerhubu, Apiary nebo podobných nástrojů), které umožní vytěžování dat ze systému Kramerius a pomocné databáze Kramerius plus.

**Základní požadovaná funkčnost:**

DL4DH Feeder

* bude obsahovat správu uživatelů a autentifikaci přes Shibboleth
* umožní extrakci, exporty a vizualizace tagovaných údajů a jejich vazeb
* podá informaci o kvalitě dat v Krameriu a kvalitě obohacení, případně o kvalitě jednotlivých verzí dat
* bude provázán na uživatelské kolekce v Krameriovi (bude jen jedna správa uživatelů pro Krameria i Feeder)
* bude podporovat formáty exportu: TEI, T/CSV, JSON, plain text, PAGE, ALTO (pokud je k dispozici v systému Kramerius) + RAW export (data v originální podobě)
* umožní vyhledávání publikací na základě geografických údajů zmíněných v textu publikace výběrem z mapy
* umožní vyhledávání na základě dotazů zohledňujících tagované prvky v textu publikace (zejména chronologie a místa)

**V rámci dostupných volných člověkodnů bude možné rozšiřovat funkčnost Feederu o další prvky. Patřit mezi ně mohou např.:**

* Napojení na více digitálních knihoven Kramerius najednou (je potřeba řešit duplicity).
* Lepší vytěžitelnost digitalizovaných dokumentů, nebo částí dokumentů, které mají encyklopedický nebo slovníkový charakter (např. rejstříky a soupisy).
* (Polo)automatická extrakce do podoby databáze/linked data.
* Identifikace knižního dekoru, prázdného místa na stránce, rukopisných přípisků.
* Rozpoznání struktury textu (sloupce), typografie apod.
* Evidence historie dotazů a porovnání, zda se ve výstupu Feederu něco změnilo.
* Možnost realizovat reprodukovatelný export – (nástroj AVOBMAT to např. řeší JSON souborem zaznamenávajícím všechna vyhledávací nastavení, tentýž soubor lze pak nahrát k zopakovaní vyhledávání).
* Citace - viz níže
	+ - Extrakce citací pro linkování (pokus o identifikaci citací v textech a jejich prolinkování na příslušný dokument/článek).
		- Integrace článkových citací, příp. umožnit uživateli, aby si mohl vytvořit citaci.
		- Možnost vyhledávání podle citace.
		- Možnost vyhledání všech citací (např. u konkrétního autora apod.).
* Extrakce dat z tabulek a grafů (identifikace stránky s tabulkou).
* Vyhledání publikací potenciálně souvisejících s vybranou pasáží textu (na základě automatické obsahové analýzy / identifikace klíčových slov). Teoreticky lze řešit i vícejazyčně - přeložit a prohledat zahraniční biblio katalogy.
* Generování výstupu a automatické předání pro API existujících externích nástrojů pro textové analýzy (např. Voyant Tools). Účelem by mělo být, aby správce systému mohl nastavit UI (knoflíky) pro uživatelský export dat do externích nástrojů. Primárně by to mělo být určeno pro uživatele, kteří nejsou schopni sami používat REST API, nebo prostě pro rychlé a snadné jednorázové užití.
* Automatické překlady zobrazovaného textu (okno s automatickým překladem).
* Síťové vizualizace pro entity nebo záznamy.
* Recommender systém (např. podle žánru, předmětového hesla, podobné ortografie)
* Vizualizace TEI XML s entitami (a další anotací – viz příklad): http://lindat.mff.cuni.cz/services/teitok/parczech-ps7-1.0/index.php?action=file&cid=xmlfiles/2013-011/2013-011-01-000.xml
* Možnost využití nástroje Grobid (https://github.com/kermitt2/grobid)
* Automatická detekce a extrakce poznámkového aparátu (užitečné zejména pro publikace historiků, kde jsou často poznámky až na konci textu).
* Wrappery pro API DH Feederu v Pythonu/R.
* Možnost otáčení stran dokumentu (viz současný klient Krameria).