národní
úložiště
šedé
literatury

Metodika přípravy dat z digitálních knihoven pro využití v digitálních humanitních vědách

Lehečka, B.
2022

Dostupný z http://www.nusl.cz/ntk/nusl-511549

Dílo je chráněno podle autorského zákona č. 121/2000 Sb.

Tento dokument byl stažen z Národního úložiště šedé literatury (NUŠL).

Datum stažení: 18.06.2024

Další dokumenty můžete najít prostřednictvím vyhledávacího rozhraní nusl.cz .

Metodika přípravy dat z digitálních knihoven pro využití v digitálních humanitních vědách

Autorský tým: Boris Lehečka, David Novák, Filip Kersch, Radim Hladík, Jarmila Bíšková, Kristýna Sekyrová, František Válek, Zdenko Vozár, Norbert Bodnár, Peter Sekan, Michaela Bežová, Petr Žabička, Martin Lhoták, Pavel Straňák

Tato metodika vznikla v rámci projektu programu na podporu aplikovaného výzkumu a vývoje národní a kulturní identity (NAKI II, Ministerstvo kultury ČR) č. DG20P02OVV002 s názvem „DL4DH – vývoj nástrojů pro efektivnější využití a vytěžování dat z digitálních knihoven k posílení výzkumu digital humanities“.

Knihovna AV ČR, v. v. i.
Moravská zemská knihovna v Brně
Národní knihovna České republiky
Praha 2022

Stručný obsah

1 Úvod

Digitalizované obrazové materiály jsou dnes obvyklou a nenahraditelnou součástí národního kulturního dědictví. Digitální knihovny obsahují velké množství textů, jež jsou pro mnohé vědní obory klíčovým zdrojem informací a často i základním předmětem výzkumu. Práce s materiály z českých digitálních knihoven se však stále v mnohém neliší od práce badatele s tištěnými publikacemi: skeny, popř. text publikace je možné prohlížet po jednotlivých stranách, volně dostupná díla lze stáhnout pouze po jednotlivých publikacích apod. To vše se děje v době, kdy počítačové metody umožňují či vyžadují práci s velkými objemy dat. Na tyto tendence reaguje mj. legislativa, v poslední době např. evropská směrnice o autorském právu.

Jedna z cest, jak vědecké komunitě, zejména z oblasti společenských a humanitních věd, nabídnout materiály, které vycházejí vstříc jejich potřebám, spočívá ve vytvoření nové úrovně služeb nad digitalizovanými dokumenty, resp. digitálními knihovnami. Mezi hlavní cíle těchto služeb patří: 1) opatřit existující materiály dalšími metadaty, která umožní jejich širší využití; 2) zajistit přívětivý uživatelský i programový přístup k velkým objemům obrazových a textových dat včetně metadat; 3) podpořit aplikaci principů FAIR (Findable, Accessible, Interoperable, Reusable; Wilkinson, 2016, srov. Morselli, 2020) při nakládání s daty a metadaty.

1.1 Cíle metodiky

Tato metodika si klade za cíl nabídnout knihovnám i dalším paměťovým institucím v České republice doporučený postup při zpřístupnění velkých objemů dat pro výzkumné účely. V současné době je z tohoto pohledu zdigitalizováno již nadkritické množství dokumentů z knihovních fondů, přičemž výsledky digitalizace jsou prezentovány v různých systémech digitálních knihoven. Při jejich zpřístupňování je třeba vždy vycházet z aktuálního znění autorského zákona, ale už nyní je možné se připravit na jeho významnou novelizaci, která implementuje směrnici Evropského parlamentu a Rady č. 2019/790 a týká se mj. vytěžování textů a dat pro vědecké účely. Metodikou doporučovaná architektura nadstavbového systému pro digitální knihovny zajistí škálovatelnost, snadnou správu i rozvoj souvisejících služeb. Představené způsoby zpracování dat, jejich obohacování i výstupní formáty vycházejí z požadavků specialistů z celé škály humanitních oborů.

1.2 Uživatelé metodiky

Metodika je určena pro dva typy uživatelů: pracovníky digitálních knihoven a badatele, kteří při svém výzkumu pracují s velkými objemy textových či obrazových dat.

Pracovníci digitálních knihoven se seznámí s aktuální a očekávanou autorskoprávní ochranou digitálních děl ve vztahu k vytěžování textů a dat pro vědecké účely a s možnostmi implementace, popř. nasazení hotového systému, který bude sloužit vědecké obci. Metodika je proto určena manažerům, technickým i obsahovým správcům, vývojářům IT a metadatovým specialistům z těchto institucí.

Badatelé z různých vědních disciplín se seznámí se svými aktuálními i plánovanými právy v oblasti vytěžování textů a dat pro účely výzkumu. Metodika popisuje principy, které je vhodné při práci s těmito materiály dodržovat, takže se hodí i pro studenty a výzkumníky, kteří se s problematikou velkých dat v digitálních knihovnách teprve seznamují. V neposlední řadě bude tento text přínosný pro badatele, kteří budou využívat nástroje DL4DH pro získání dat pro výzkum. Metodika v závěru obsahuje několik námětů na případové studie z humanitněvědních oborů (religionistika, archeologie, sociologie, knihověda, versologie), které slouží jako ilustrativní ukázky, kterými se mohou inspirovat badatelé z dalších humanitních, ale i přírodovědných oborů.

1.3 Popis metodiky

Metodika se v počáteční kapitole věnuje změnám ve výzkumu v humanitních vědách, na nichž se významnou měrou podílí vývoj výpočetní techniky a nových analytických metod. Na popis principů FAIR, které je vhodné při výzkumu digitálněhumanitními metodami dodržovat, navazuje charakteristika základních textových pramenů pro výzkum v těchto oborech (korpusů a digitálních knihoven) a popis vybraných digitálních knihoven v zahraničí a v Česku, které mají v tomto ohledu badatelům co nabídnout. Podrobněji je vymezena aktuální i plánovaná česká právní úprava v oblasti vytěžování textů a dat pro účely výzkumu.

Představení projektu DL4DH začíná popisem standardů digitalizace definovaných Národní knihovnou ČR a seznámením se systémem Kramerius, který využívá většina digitálních knihoven v Česku. V závěru jsou popsány hlavní principy projektu DL4DH.

Část popisující předzpracování dat a metadat v rámci projektu DL4DH je doplněna o ilustrativní výčet dalších nástrojů pro (před)zpracování dat, které je možné při vědecké práci využít.

Na to navazuje popis architektury implementovaného systému. Další kapitola informuje obecně o nakládání s různými typy dat při vědecké práci a přibližuje formáty uložení obrazových a textových dat včetně metadat v projektu DL4H i o způsobu, jak údaje ze systému získat. Následuje kapitola, která shrnuje základy práce s vyvinutými nástroji z pohledu uživatele, tj. správce digitální knihovny a badatele. Popisuje jak uživatelské rozhraní ve webovém prohlížeči, tak základní principy ovládání pomocí programového rozhraní REST API. V závěru jsou zmíněna známá omezení a problémy, s nimiž je nutné při práci nástroji DL4DH počítat.

Závěr metodiky je věnován námětům na případové studie, které na výzkumných tématech z různých oborů ilustrují využití implementovaných nástrojů DL4DH.

1.4 Srovnání novosti postupů

Velká část digitálních knihoven v Česku využívá informační systém Kramerius, jehož primárním cílem je nabídnout uživatelům digitalizované publikace v jejich obrazové podobě. Textová vrstva (ve formě prostého, jenom částečně strukturovaného a neformátovaného textu) je dostupná pouze pro jednotlivé strany, případně ve větších celcích po stažení dokumentu ve formátu PDF.

Knihovny provozované v systému Kramerius nabízejí REST API pro přístup k metadatům o digitálních dokumentech, popř. k jejich obrazům. Toto rozhraní je ale navrženo tak, aby pracovalo s metadaty o celé publikaci, popř. s jednotlivými stranami. Pro práci s kolekcemi dokumentů a jejich stahování v různých formátech se tento systém nehodí.

Programové rozhraní API pro stahování velkých objemů dat je možné najít u agregátorů dat (např. poskytovatelů statistických údajů) nebo v rámci komerčních služeb typu Twitter apod. Knihovní instituce v Česku obdobné služby dosud nenabízejí, zahraniční digitální knihovny umožňují tímto způsobem přistupovat k metadatům.[1] To je dáno zejména aktuálním pojetím ochrany autorských práv, v zemích Evropské unie pak teprve postupným pronikám směrnice Evropského parlamentu a Rady č. 2019/790 do národních právních norem. Podle této úpravy do autorských práv nezasahuje ten, kdo vytěžuje texty a data pro vědecké účely.

Svými službami, funkcemi uživatelského rozhraní a částečně i programovým rozhraním má k řešením představeným v metodice blízko společný projekt zhruba 150 zahraničních knihoven nazvaný *HathiTrust*. Badatelé mohou pracovat s kolekcemi dokumentů, ale kvůli autorským

[1] Viz např. přehled na https://www.programmableweb.com/category/library/api.

právům mají přístup pouze k agregovaným datům z těchto děl (např. k frekvenci slovních tvarů na jednotlivých stranách). Přístup k plným textům je možný pouze ve virtuálním počítači, který provozuje *HathiTrust Research Center*. Předkládaná metodika je proto první pokus o systematické uchopení této problematiky v českém prostředí.

2 Digitální humanitní vědy

Stav lidského poznání se proměňuje na základě postupného zpřesňování poznatků o světě kolem nás i v nás. Generace badatelů navazují na své předchůdce, verifikují jejich poznatky a přicházejí s novými teoriemi, pro které hledají oporu ve studované materii, tedy ve vědeckých datech. Podobné proměně podléhají i nástroje, které se k řešení výzkumných otázek používají. Výraznou změnu pro vědce přinesl vývoj počítačové techniky, která umožnila explicitnější a přesnější zacházení s daty a zároveň dovolila digitalizovat a zpřístupnit kulturní dědictví širokému počtu badatelů i laické veřejnosti. V posledním období lze díky nárůstu úložných kapacit a výpočetního výkonu pracovat s mnohem většími objemy dat než v minulosti. Všechny tyto faktory přispěly k tomu, že badatelé mohou při práci s daty využívat nové metody, jako jsou např. strojové učení a umělá inteligence. Nové metody a předměty studia, které jsou reakcí na tyto změny, daly vzniknout multivědnímu oboru, který se nazývá *digitální humanitní vědy* nebo také *digital humanities* (dále též DH; Luhmann, 2022).

Přírodní a humanitní vědy již delší dobu sdílejí data napříč obory a systematicky budují datové repozitáře, na které jsou navázány služby jejich základní analýzy. Tato úložiště, původně vytvářená jako izolované výstupy jednotlivých výzkumných projektů, se postupně propojují napříč pracovišti a obory, což zvyšuje efektivitu jejich využití a umožňuje sdílet výstupy v celoevropském i celosvětovém kontextu. To vede k rozšíření efektivity i záběru bádání a samozřejmě i k otevírání nových otázek. Specifikem humanitních věd je množství odborných i obecných textů, které jsou obsaženy v digitálních knihovnách, jež jsou pro určité obory (například obecnou i matematickou lingvistiku, dějiny literatury, sociologii, dějiny knižní kultury atd.) klíčovým zdrojem informací a mnohdy základním předmětem výzkumu. Pro další vědní obory jde o podstatný doplňkový zdroj informací a souhrnně jde o nedílnou součást kulturního dědictví. Přestože se jednotlivé obory humanitních věd technologicky rozvíjejí, hledání zdrojů pro analytickou práci za využití digitálních nástrojů je stále velmi složité (srov. Lhoták, 2020).

Digitální dokumenty mají i mnoho dalších výhod: jsou např. snadno šiřitelné, mohou být dostupné 24 hodin denně, umožňují prohledávání a porovnávání plných textů digitalizovaných dokumentů. Tyto vlastnosti a nástroje činí z digitálních knihoven nový a mimořádně bohatý zdroj dat pro výzkum. Digitalizace starší vydavatelské produkce vrací do současnosti pozapomenuté texty a umožňuje aplikovat zcela nové postupy a vazby mezi poznatky, které práce s fyzickým knihovním fondem neumožňovala.

2.1 Práce s digitálními daty

Výzkumná data představují jeden z hlavních pilířů, díky nimž se může bádání efektivně posouvat vpřed. Při práci s nimi je vhodné mít na zřeteli doporučení, která se v souvislosti s novými metodami badatelské práce ustálila (principy FAIR), ale i aktuálně platný právní rámec, který umožňuje využívat digitálně dostupné publikace k různým účelům, nebo jejich využití omezuje.

2.1.1 Principy FAIR

Pro publikování výzkumných dat a metadat vznikl soubor metodických pokynů označovaný zkratkou FAIR (z anglických výrazů Findable, Accessible, Interoperable, Reusable; Morselli, 2020; GO FAIR, 2022). Tyto principy počítají s využitím dat nejen uživatelem-vědcem, ale také prostřednictvím počítačového zpracování.

FAIR data by měla být dohledatelná, a to díky přiřazenému jedinečnému perzistentnímu identifikátoru (PID; v praxi např. DOI, či Handle) a dostatečným metadatům, která jsou prohledávatelná pomocí veřejně dostupného systému. Data i metadata musí být přístupná pomocí standardních komunikačních protokolů, které v případě potřeby umožňují ověřit totožnost uživatele a jeho práva k (meta)datům (pomocí autentizace a autorizace). Interoperabilita dat je možná díky tomu, že se pro reprezentaci znalostí (v datech a metadatech) používá formální, dostupný, sdílený a široce aplikovatelný jazyk (ideálně umožňující strojové čtení), ale i díky vzájemným odkazům mezi (meta)daty. Znovuvyužitelnost (meta)dat zaručuje dostatečné množství popisných vlastností s adekvátními hodnotami, dodržování standardů vědecké komunity pro daný obor a rovněž publikování s uvedením zdroje a pod jasnou licencí.

Aplikace principů FAIR je podstatná pro všechny vědní oblasti, avšak právě z pohledu humanitních věd jde o revolučně nový přístup. Samotné pojetí pramenů jako zdrojů dat (nikoli jako prostých informačních médií) se více prosazuje až s nástupem DH. Principy FAIR nabízejí návod, jak se posunout k co nejefektivnější, nejdůvěryhodnější a nejvíce replikovatelné vědecké práci, tedy jak dosáhnout kvalit, které bývají v humanitních vědách opomíjeny či cíleně přehlíženy. Tyto přístupy se velice dobře doplňují s vývojem v oblasti digitálních korpusů a knihoven s jejich potřebami.

2.2 Od opakovaně využitelných dat k reprodukovatelnému výzkumu

Opakovatelnost postupů a dosahování totožných výsledků patří k principům vědeckého poznání od počátku moderní vědy. Rozvoj digitálních metod vedle tradičního požadavku na replikovatelnost výzkumu (stejná metoda vede na nových datech ke stejným výsledkům) akcentuje také zásadu reprodukovatelnosti, tj. ověření integrity výzkumu opakováním metodických postupů na původních datech (National Academies of Sciences, Engineering, 2019). Programový kód využívá komputační reprodukovatelnost jako formální pojítko mezi surovými daty a publikovanými výsledky bádání. Přestože relevance těchto principů pro humanitní obory bývá s ohledem na jejich idiografický charakter zpochybňována (Penders, 2019), empiricky založené humanitní výzkumné otázky zkoumané s využitím počítačových postupů mohou být kritériem reprodukovatelnosti poměřovány (Peels, 2019). DL4DH podporuje reprodukovatelnost výzkumu tím, že badatelům poskytuje data v souladu s principy FAIR, eviduje u zpracovaných dokumentů časové značky a uchovává historii dotazů do databáze Kramerius+. Díky těmto opatřením je možné opakovaně vytvářet totožné datové sady nebo identifikovat změny v datových podkladech, které mohou mít dopad na výsledky po spuštění totožného analytického kódu.

V oblasti humanitních věd se postupně utvářely různé typy textových zdrojů, které sloužily k dalšímu výzkumu. Ty hlavní lze rozčlenit na jazykové korpusy a digitální knihovny.

2.3 Jazykové korpusy

Jazykové korpusy představují soubor textů určitého jazyka. Jedná se o „rozsáhlý soubor autentických textů (psaných nebo mluvených) převedený do elektronické podoby v jednotném formátu tak, aby v něm bylo možné jednoduše vyhledávat jazykové jevy“[2]. Často slouží jako lexikologický a lexikografický nástroj, ale používají se i v jiných oblastech, které využívají texty jako zdroje poznání reality (historie, sociologie, psychologie apod.). Na rozdíl od digitálních knihoven korpusy často nemají přímou vazbu na digitalizát, z něhož vycházejí[3]: v některých případech jde o tzv. born-digital texty bez klasické tištěné předlohy (nebo tato vazba není zaznamenána, např. u elektronické verze periodik). Zpočátku vznikaly textové korpusy na základě tištěných předloh, v poslední době vznikají také korpusy sestavené z textů dostupných na

[2] https://wiki.korpus.cz/doku.php/pojmy:korpus
[3] Tj. nelze jednoduše zobrazit nebo přejít na odpovídající odstavec či stránku originálního dokumentu.

webových stránkách (viz např. projekt Aranea[4]). V současnosti existuje velké množství korpusů různých národních jazyků s obecným i specifickým zaměřením na určité oblasti jazyka (poezie, historické texty, parlamentní projevy apod.).

V českém prostředí hrají na poli jazykových korpusů významnou roli velké výzkumné infrastruktury Český národní korpus (ČNK) a LINDAT/CLARIAH-CZ, které se věnují nejen přípravě synchronních a diachronních korpusů, ale i vývoji nástrojů pro jejich analýzu, jako je např. morfologický analyzátor nebo korpusový manažer. ČNK vytváří s odstupem pěti let stomilionové žánrově vyvážené korpusy, které jsou od roku 2014 referenční, tj. neměnné. Největší sbírku starších českých textů nabízí v současné době Staro- a středněčeská textová banka[5].

2.4 Digitální knihovny

V oblasti archivace kulturního dědictví sehrála významnou roli digitalizace knihovních fondů. S rozvojem technologií pro optické rozpoznávání znaků (OCR) a v poslední době i ručně psaného textu (HTR) mohly vzniknout digitální knihovny, které vedle naskenovaných obrazů obsahují také plné texty těchto publikací. Rozdíl proti jazykovým korpusům spočívá v tom, že tyto texty nebývají strukturované, respektive strukturace existuje pouze na základní úrovni (strany a odstavce), a to i vzhledem k velkému množství typů publikací. Digitalizáty jsou propojené s bibliografickými záznamy, které vznikají v informačních systémech jednotlivých knihoven.[6] Ze zahraničí je např. známý projekt *Google Books*[7] (v jehož rámci se digitalizovaly i publikace z českých knihoven), méně rozsáhlé fondy jsou k dispozici díky projektům, jako jsou *English Books Online Text*[8] nebo *HathiTrust Digital Library*[9].

Digitalizace v Česku byla od svého počátku využívána především jako jeden z hlavních nástrojů dlouhodobé ochrany tištěných i psaných dokumentů, protože papír jako organická hmota postupně stárne a rozpadá se. Analogovým fondům nadto hrozí i další rizika jejich poškození či zničení: přírodní katastrofy, požáry, válečné události, povodně, uchovávání v nesprávných podmínkách apod. Knihovní fondy českých knihoven se digitalizují od poloviny 90. let minulého století, zpočátku pomocí mikrosnímkování na mikrofilmy a mikrofiše, které byly později digitalizovány. Na konci 90. let se začalo digitalizovat i tzv. hybridním způsobem, tj. nejdříve

[4] http://ucts.uniba.sk/aranea_about/index.html
[5] https://korpus.vokabular.ujc.cas.cz/first_form
[6] V některých případech, např. v projektu Manuscriptorium, vznikají bibliografické záznamy také digitalizací starších knižních katalogů, takže uváděné údaje nemusejí být aktuální.
[7] https://books.google.cz
[8] https://textcreationpartnership.org/tcp-texts/eebo-tcp-early-english-books-online/
[9] https://www.hathitrust.org

se titul převedl na mikrofilm a posléze se digitalizoval na speciálních skenerech, aby mohl být dále zpřístupněn v digitální podobě. V současné době (stav k 30. 4. 2022) je digitalizováno přes 0,5 mil. dokumentů, z toho přes 300 tis. knih, 160 tis. rukopisů a starých tisků, 60 tis. svazků/ročníků periodik (pokračujících zdrojů) a desetitisíce map, grafik, hudebnin, a dalších druhů dokumentů včetně několika tisíc zvukových záznamů.[10] Digitalizují se především bohemikální dokumenty, což zahrnuje i cizojazyčné publikace, které jsou s českými zeměmi a jejich knižní produkcí nějakým způsobem provázané. Celkově se odhaduje, že bylo v Českých knihovnách již digitalizováno přes 80 milionů stran.

Následující kapitoly na vybraných příkladech popisují aktuální podobu digitálních knihoven v zahraničí a Česku. V přehledu jsou uvedeny aplikace s výraznou přidanou hodnotou pro badatele z oblasti digitálních humanitních věd.

2.4.1 Zahraniční digitální knihovny

Řada digitálních knihoven provozovaných národními knihovnami nebo jinými významnými institucemi nabízí přístup ke svým datům a metadatům pro výzkumné účely. Odkazy na rozhraní pro přístup k sadám dat, API nebo souvisejícím aplikacím bývají obvykle seskupeny na stránkách označených jako "laboratoře" dané knihovny. Lze zmínit například laboratoře americké *Library of Congress*[11], dánské *Det Kgl. Bibliotek*[12], nizozemské *Koninklijke Bibliotheek*[13], britské *British Library*[14], rakouské *Österreichische Nationalbibliothek*[15] nebo *National Library of Australia* v rámci digitální knihovny Trove[16]. Za podrobnější zmínku stojí služby nabízené sdružením HathiTrust a *Bavorskou zemskou knihovnou*, které nabízejí nejen přístup k obrazovým datům, ale i k plným textům volných děl (HathiTrust) a pro výzkumné účely i děl chráněných autorskými zákony (Bavorská zemská knihovna). HathiTrust poskytuje také agregovaná data z dokumentů a programové knihovny pro práci s nimi.

[10] Aktuální údaje jsou dostupné v Registru digitalizace (2017), který eviduje stav digitalizace fondů v Česku.
[11] https://labs.loc.gov
[12] https://labs.kb.dk
[13] https://lab.kb.nl
[14] https://www.bl.uk/projects/british-library-labs
[15] https://labs.onb.ac.at/en/
[16] https://trove.nla.gov.au/landing/research

Projekt HathiTrust sdružuje zejména knihovny z anglojazyčných zemí. Průběžně digitalizované publikace se stávají součástí jedné digitální knihovny, pojmenované *HathiTrust Digital Library*. V roce 2022 šlo přibližně o 17,1 mil. publikací, z toho 6,57 mil. bylo volně dostupných a 10,55 mil. bylo chráněno aktuálně platnými autorskými zákony. Díla lze prohledávat na základě bibliografických metadat nebo v rámci automaticky rozpoznaného textu. Uživatelé mohou vytvářet a případně sdílet s ostatními uživateli vlastní kolekce dokumentů. Pokud nejsou díla chráněna autorským právem, lze je prohlížet po jednotlivých stranách ve formě digitálních obrázků nebo prostého nestrukturovaného textu. Volně dostupná díla je dále možné stáhnout ve formátu PDF s textovou vrstvou, např. prostřednictvím *Google Books* (v tomto případě jsou však obrázky černobílé, nikoli plně barevné).

Uživatelé z partnerských institucí projektu mají v rámci *HathiTrust Research Center* k dispozici nástroje pro další analýzu zvolené sady publikací. Jedná se o rozpoznání entit[17], tzv. „extrahované prvky“[18], identifikace témat metodou InPhO[19], počty tokenů (včetně morfologické analýzy) a slovní mraky[20]. Tato metadata obsahují sumarizované údaje pro jednotlivé strany, resp. kompletní publikace, čímž je zaručeno, že nedochází k porušování autorských práv. Některé druhy sumarizačních výstupů se pravidelně aktualizují a jsou k dispozici všem zájemcům (např. geografická jména v anglojazyčné literatuře z let 1701–2011, Wilkens, 2020). Badatelé mají k dispozici programové knihovny v jazyce Python pro práci s vytvořenými kolekcemi metadat, případně samostatný virtuální stroj s nezbytným softwarem a přístupem ke všem textovým zdrojům HathiTrust.

Digitální sbírka Bavorské zemské knihovny čítá přibližně 2,7 mil. digitalizovaných publikací z historických fondů. Díla lze prohledávat na základě bibliografických metadat nebo v rámci automaticky rozpoznaného textu. V druhém případě se ve výsledcích vyhledávání zobrazuje detail nalezené stránky včetně zvýrazněného hledaného slova. Při listování publikací je k dispozici několik funkcí pro ovládání naskenovaného obrázku: vedle otočení obrazu jde o nastavení jeho jasu a kontrastu, změnu barev na odstíny šedé nebo inverzní barvy. Díky technologii IIIF může uživatel generovat a sdílet odkaz nejen na jednotlivé stránky, ale i jejich výřezy. Nad obrazovou vrstvou je možné zobrazit vrstvu s textem, u níž lze ovládat barvu a průhlednost. Obraz je dostupný ke stažení ve formátu JPEG v různých rozlišeních, nebo ve formátu PDF.

[17] https://analytics.hathitrust.org/algorithms/Named_Entity_Recognizer
[18] https://wiki.htrc.illinois.edu/display/COM/HTRC+Derived+Datasets
[19] https://inpho.github.io/topic-explorer/
[20] https://analytics.hathitrust.org/algorithms/Token_Count_and_Tag_Cloud_Creator

Chce-li uživatel publikaci využít pro badatelské účely (v takovém případě je vyžadován souhlas prostřednictvím e-mailu), obsahují stahovaná data obrázky a samostatné soubory s textem jednotlivých stran, v těchto případech může zpracování trvat až čtyři týdny.

2.4.2 Digitální knihovny v České republice

V Česku probíhá digitalizace publikací ve velkém měřítku zejména ve veřejných knihovnách, zřizovaných nebo provozovaných státem, kraji či obcemi, a v Knihovně Akademie věd. Dále se na digitalizaci podílejí specializované knihovny (Národní lékařská knihovna ap.) umístěné ve veřejných objektech, vysokoškolské knihovny nebo muzejní, archivní a galerijní knihovny. Digitální knihovny, které toto bohatství zpřístupňují, jsou obvykle vázány na jednotlivé instituce, ale v rámci společné koordinace sil a prostředků vznikly dva hlavní informační systémy, které se zaměřují na různé typy publikací. Systém *Kramerius* je primárně určen pro digitalizované knihovní sbírky, monografie a periodika. Naproti tomu *Manuscriptorium* umožňuje snadný přístup k soustředěným informacím o historických fondech.

Digitální knihovny v českém prostředí používají ke zpřístupňování digitalizovaných publikací především opensourcový systém Kramerius (2022). Rozcestník vybraných knihoven s tímto systémem je pojmenován Digitální knihovna (2022) a je k dispozici na adrese https://www.digitalniknihovna.cz. Uživatel může publikace vyhledat podle bibliografických dat i na základě fulltextového hledání. Následně lze nalezeným dokumentem listovat po jednotlivých digitalizovaných stranách, zobrazit rozpoznaný text pro aktuální stranu (je-li k dispozici), případně stáhnout publikaci ve formátu JPEG nebo PDF, umožňuje-li to aktuálně platná licence k dílu. Podrobnější popis Krameria je uveden v kapitole 3.3.

Specifické postavení v infrastruktuře instalovaných systémů Kramerius má Česká digitální knihovna (2022). Jedná se o agregátor mnoha českých Krameriů, jehož hlavním přínosem je prohledávání a prohlížení publikací ze všech zapojených Krameriů na jednom místě. Česká digitální knihovna je také národním agregátorem pro portál *Europeana*.[21]

Digitální knihovna *Manuscriptorium* je certifikovaným doménovým subagregátorem projektu *Europeana* pro oblast historických fondů, do níž je zapojeno více než 180 institucí z Evropy. Jedná se o největší digitální knihovnu zaměřenou na starší písemné dědictví na světě, která obsahuje více než 154 tis. digitálních dokumentů různého typu (rukopisy, inkunábule, staré tisky,

[21] https://www.europeana.eu/cs

historické mapy aj.) a více než 230 tis. katalogových záznamů. Vzhledem k povaze digitalizovaných pramenů (převažují rukopisy a staré tisky) jsou k dispozici plné texty jen v omezené míře.

Poslední, čtvrtá verze webové aplikace Manuscritporium z roku 2021 umožnila přechod k virtuálnímu badatelskému prostředí, které např. zajišťuje práci s dokumenty kompatibilními s technologiemi IIIF z libovolného externího zdroje, a to díky integraci prohlížeče Mirador. Data každého digitalizátu jsou popsána dle schéma TEI P5 ENRICH[22], v rámci digitální knihovny má digitalizát přidělený identifikátor URI a jeho použití podléhá různým druhům licence Creative Commons (CC). Aktuální prostředí umožňuje přihlášeným badatelům vytvářet vlastní kolekce dokumentů nebo jejich částí a přidávat k jednotlivým položkám poznámky.

2.5 Data z digitálních knihoven a český autorský zákon

Drtivá většina materiálů, které české knihovny uchovávají a digitalizují, mají povahu díla, jež je výsledkem jedinečné tvůrčí činnosti fyzické osoby (autora). Proto se na ně vztahuje ochrana autorských práv, která se v českém právním prostředí řídí autorským zákonem č. 121/2000 Sb. (zkráceně AZ).

2.5.1 Aktuální právní úprava

Digitální knihovny se při zpřístupňování vlastněných děl řídí povinnostmi, které jim AZ v aktuálně platném znění ukládá. Pro zveřejnění digitalizovaného dokumentu, u kterého neuplynulo 70 let od smrti všech jeho autorů, je vždy vyžadován souhlas autora nebo držitele majetkových práv.[23] Výjimku z tohoto pravidla tvoří úřední díla (právní předpisy, veřejné listiny, veřejně přístupné rejstříky, obecní kroniky apod.), které lze z důvodu veřejného zájmu zpřístupňovat bez omezení. V současné době je prostřednictvím digitálních knihoven se systémem Kramerius na internetu volně přístupno přibližně 15 % z toho, co bylo dosud v Česku digitalizováno (viz kapitola 3.1.1). Přístup k chráněným dílům je omezen: uživatelé mohou digitální kopii užívat

[22] Jelikož však dokumenty s kořenovým elementem TEI neobsahují označení jmenného prostoru, neodpovídají dokumenty z Manuscriptoria deklarovanému standardu.

[23] AZ se týká pouze děl, jejichž autoři jsou občany České republiky, ať už byla díla uveřejněna kdekoli. U občanů jiných států se autorské právo řídí mezinárodními smlouvami, zejména tzv. Bernskou úmluvou, evropskými předpisy sjednocujícími autorské právo a relevantními národními předpisy, které jsou ale na poli autorského práva zvlášť sjednocené (https://wipolex.wipo.int/en/text/283698; https://cs.wikisource.org/wiki/Bernská_úmluva_o_ochraně_literárních_a_uměleckých_děl), která vyžaduje délku ochrany minimálně 50 let po smrti autora; většina států světa (včetně Evropské unie a USA) tuto ochranu stanovuje na 70 let.

pouze za specifických podmínek definovaných zákonem, obvykle v prostorách knihovny, která dokument digitalizovala.

Novelizace AZ v roce 2017 zavedla pojem „díla nedostupná na trhu“ (zkráceně DNNT) a přinesla nové možnosti licencování služeb knihoven ze strany kolektivních správců. Tato úprava rozšířila oblasti, v nichž je možno nově vykonávat tzv. rozšířenou kolektivní správu usnadňující licencování některých způsobů užití děl knihovnami, které působí podle knihovního zákona č. 257/2001 Sb.

DNNT jsou vymezena autorským zákonem (§ 97f AZ) jako: „díla slovesná, včetně děl do nich vložených nebo začleněných nebo tvořících jejich nedílnou součást (tedy ilustrací apod.), pro která byl autorem (nositelem práv), knihovnou nebo kolektivním správcem podán návrh na zařazení do seznamu děl nedostupných na trhu uveřejněného na webu Národní knihovny, a jejichž užití není zjevně předmětem prodejních nebo licenčních podmínek, které vylučují možnost označit je za nedostupná na trhu (například možnost zakoupení e-knihy), a která nebylo možno v druhově shodném nebo obdobném vyjádření ve lhůtě 6 měsíců od podání návrhu, při vynaložení přiměřeného úsilí a za obvyklých podmínek, opatřit za úplatu v běžné obchodní síti, a která byla na tomto základě Národní knihovnou zařazena do seznamu děl nedostupných na trhu, přičemž autor (nositel práv) nevyzval Národní knihovnu k vyřazení svého díla, ať z návrhů nebo ze seznamu děl nedostupných na trhu.“[24]

Seznam děl nedostupných na trhu je zveřejněn na webových stránkách Národní knihovny ČR (Seznam děl nedostupných na trhu, 2021) a umožňuje držitelům práv, aby rozhodli o zpřístupnění, nebo o zákazu přístupu ke svým dílům.

V roce 2019 dalo Ministerstvo kultury souhlas k uzavření licenčních smluv mezi Národní knihovnou ČR a kolektivními správci DILIA a OOA-S. Smlouva je uzavřena na období pěti let do roku 2023 a umožňuje zpřístupnění digitalizovaných monografií vydaných na území České republiky před více než 20 lety a periodik vydaných před více než 10 lety. Jedná se o dokumenty, které jsou chráněny autorským zákonem (tj. neuplynulo 70 let od smrti autora) a které zároveň nejsou dostupné na trhu. Zobrazení děl je možné pomocí vzdáleného přístupu (prostřednictvím systému Kramerius) nebo terminálu (počítače umístěného v knihovně). V druhém případě má uživatel navíc přístup k dílům vydaným mezi lety 2001 až 2007.

[24] Viz https://www.zakonyprolidi.cz/cs/2000-121#f6026806

Podle aktuálního znění autorského zákona je možné v rámci digitálních knihoven poskytovat volná díla k libovolnému užití, kdy uživatel neužívá dílo dehonestujícím způsobem nebo si nenárokuje autorství díla. Metadata a paradata (viz kapitola 3.2), pokud netvoří databázi, nejsou považována za autorské dílo, takže je lze rovněž poskytovat bez omezení. V případě, že metadata a/nebo paradata databázi tvoří, je k jejímu poskytnutí potřeba souhlas pořizovatele databáze, tj. knihovny nebo instituce, která tyto údaje vytvořila. Na všechny tyto údaje mají nárok i anonymní uživatelé bez ověřené identity.

Zveřejnění volných děl nemůže být omezeno autorským právem. Instituce se ale při volbě licence pro zveřejňovaná volná díla mohou řídit strategickým, politickým nebo obchodním rozhodnutím, vždy by však měly postupovat transparentně a v souladu s platným právním řádem.

Specifická situace nastává při obohacování textových dat (anotace) pomocí externích nástrojů nebo služeb. Na obohacující údaje lze v tomto případě pohlížet jako na databázi. Pokud by knihovna anotaci objednala za úplatu na základě jasných parametrů nebo zaplatila za software, který anotaci provede, stala by se pořizovatelem databáze a mohla by zvolit licenci takto vytvořených dat. Využívá-li knihovna software nebo služby bezplatně, je databáze chráněna jen autorským právem (a ne právem pořizovatele), takže se při zveřejnění obohacených dat musí řídit licenčními podmínkami a doporučeními subjektu, který aplikaci nebo službu poskytuje.

Z praktického hlediska a z pohledu dodržování principů FAIR je vhodné opatřit všechny poskytované údaje konkrétní licencí, u volných děl např. CC BY[25]. Licence k dílu bude uvedena v rámci metadat k jednotlivým dílům, popř. k databázi. Totéž se týká informací o nástrojích, které byly využity při obohacování dat a metadat jednotlivých děl (viz kapitola 4). Systém, který umožní údaje stahovat, by měl uživatele o licencích informovat souhrnně při stahování dat, pokud využije webové uživatelské rozhraní; při programovém stahování dat pomocí REST API postačí fakt, že konkrétní licence budou součástí stažených metadat.

2.5.2 Směrnice EU o autorském právu na jednotném digitálním trhu

Před třemi lety byla schválena Směrnice Evropského parlamentu a Rady (EU) 2019/790 ze dne 17. dubna 2019 o autorském právu a právech s ním souvisejících na jednotném digitálním trhu a o změně směrnic 96/9/ES a 2001/29/ES[26]. Její implementaci do českého právního systému

[25] https://creativecommons.org/licenses/by/4.0/
[26] https://eur-lex.europa.eu/legal-content/CS/TXT/?uri=CELEX:32019L0790

představuje dosud neschválená úprava AZ. Bude-li schválen Sněmovní tisk 31/0[27] v navrhovaném znění, bude mít vliv na vytěžování textů a dat za účelem získávání nových poznatků a objevování nových trendů.

Do autorského práva podle navrhovaného § 39d nezasahuje vysoká škola, která jako součást své činnosti provádí vědecký výzkum, nebo právnická osoba, jejímž hlavním cílem je provádět vědecký výzkum nebo vykonávat vzdělávací činnost zahrnující rovněž vědecký výzkum, jestliže je vědecký výzkum této vysoké školy nebo právnické osoby prováděn tak, aby přístup k jeho výsledkům nebyl přednostně umožněn tomu, kdo na tuto vysokou školu nebo právnickou osobu vykonává rozhodující vliv, a současně tak, aby výzkum byl prováděn ve veřejném zájmu nebo na neziskovém základě nebo tak, že všechny zisky jsou zpětně investovány do vědeckého výzkumu této vysoké školy nebo právnické osoby.

Podmínky využívaní děl chráněných AZ výše uvedeným způsobem jsou tedy následující:

a) musí jít o vysokou školu (včetně jejích knihoven), výzkumný ústav nebo jakýkoli jiný subjekt, jejichž hlavním cílem je provádět vědecký výzkum nebo vykonávat vzdělávací činnosti, jejichž součástí je rovněž vědecký výzkum;

subjekt využívá díla:

b) nekomerčně nebo tak, že zpětně investuje všechny zisky do svého vědeckého výzkumu,
c) v souladu s úkoly ve veřejném zájmu uznávanými členským státem,
d) zároveň takovým způsobem, že přístup k výsledkům tohoto vědeckého výzkumu není přednostně umožněn podniku, který na tuto organizaci vykonává rozhodující vliv (např. prostřednictvím propojených podniků).

Ve výsledku to prakticky znamená, že vědci, kteří mají zákonný přístup k dílu nebo nahrávce v elektronickém formátu (například prostřednictvím knihovny vlastní instituce nebo veřejné knihovny), mohou volně pořizovat další kopie těchto děl nebo nahrávek, aby mohli provádět datovou analýzu jejich obsahu, aniž by museli požádat o povolení vlastníka autorských práv (například vydavatele nebo nahrávací společnost). To platí bez ohledu na podmínky stanovené v jakékoli licenční smlouvě mezi vydavatelem a knihovnou. Výzkum však musí být nekomerční povahy a výzkumník musí uvést zdroj, pokud to není z praktických důvodů nemožné.

[27] https://www.psp.cz/sqw/text/tiskt.sqw?O=9&CT=31&CT1=0

Jelikož se výjimka na užití děl vztahuje na instituce, potažmo její zaměstnance nebo studenty, jeví se jako praktické ověřovat v rámci řešení DL4DH identitu uživatele pomocí přihlášení prostřednictvím České akademické federace identit eduID.cz (2021), přičemž ověření bude povoleno pouze pracovníkům a studentům z institucí uvedených v Rejstříku veřejných výzkumných institucí[28].

[28] https://www.msmt.cz/vyzkum-a-vyvoj-2/rejstrik-verejnych-vyzkumnych-instituci-1

3 Digital Libraries for Digital Humanities (DL4DH)

Velkou výzvou poslední doby v oblasti DH představuje práce s velkými objemy dat (Kaplan, 2015), které umožňují zkoumat kulturní, sociální, jazykové a další jevy, které se v menších datových souborech neprojeví. Pro analýzu jazyka existovaly relativně rozsáhlé jazykové korpusy, které se však hodí zejména k lingvistickým výzkumům. Proto badatelé z různých oborů volali po zpřístupnění takového zdroje dat, který by bylo možné využít pro výzkum diachronní i synchronní, a to nejen na základě dat připravených primárně s ohledem na jazykový výzkum.

Potřeba reagovat na nové metody a předměty studia v oblasti humanitních věd vedla po vzoru zahraničních sdružení k založení České asociace pro digitální humanitní vědy, z. s. (CzADH)[29], která „je spolkem studentů, vědeckých pracovníků a dalších zainteresovaných osob, jejichž cílem je napomáhat rozvoji digitálních humanitních věd ve školských a vědeckých zařízeních, ve spolcích a mezi širší veřejností“.

Na sílící potřebu přístupu k novým zdrojům a vědeckým metodám reagoval i projekt DL4DH, zaměřený na vytěžování obsahu digitálních knihoven. Na jeho řešení se v letech 2020–2022 podílely Knihovna Akademie věd ČR (KNAV), Národní knihovna ČR (NK) a Moravská zemská knihovna v Brně (MZK), a to ve spolupráci s dalšími odborníky z oblasti humanitních věd, kteří se z velké části angažují také ve spolku CzADH. Jedním z výsledků uvedeného projektu je i tato metodika.

Humanitně zaměření vědečtí pracovníci potřebují jednak data obohacená tak, aby se zefektivnilo vyhledávání relevantních zdrojů, jednak odpovídající prostředky pro jejich vytěžení, díky čemuž se usnadní důležitá část badatelské práce. Řešitelé projektu se proto shodli na těchto základních vlastnostech vyvíjené platformy:

- bude vybudována nad systémem Kramerius, který představuje nejlepší praxi v oblasti českých digitálních knihoven;
- zprostředkuje badatelům přístup k velkým objemům dat (nikoli po jednotlivých publikacích);
- umožní přístup k datům a metadatům pomocí programového rozhraní (REST API);
- dovolí sdílet vyhledaná (meta)data;
- zpřístupní údaje ve standardizovaných formátech.

[29] https://www.czadh.cz

Následující kapitoly popisují principy digitalizace rukopisné, tištěné i digitální produkce v českém prostředí a digitální knihovnu Kramerius, která slouží k jejich zpřístupnění. Na ně navazuje popis nástrojů DL4DH, které vznikly pro potřeby badatelské práce v oblasti digitálních humanitních věd, mj. pro práci s velkými objemy dat.

3.1 Instituce a jejich digitální knihovny

Digitální knihovny vznikají digitalizací knižních, mapových a dalších fondů jednotlivých institucí; může jít o samostatné knihovny nebo o knihovny v rámci jiných institucí (archivu, muzea, vysoké školy, galerie apod.). Digitalizované fondy pak tyto instituce mohou zpřístupňovat svým čtenářům v souladu s autorským zákonem. Proto je potřeba počítat s tím, že v systému Kramerius, který provozuje konkrétní instituce, budou obsažena zejména díla z jejího knihovního fondu.

Na základě dohod mezi jednotlivými knihovnami mohou být identické digitální kopie (včetně metadat) součástí několika digitálních knihoven[30]. K této replikaci dochází pouze v případech, kdy se jedná o dokument, který mají ve svém fyzickém fondu knihovny, mezi nimiž replikace probíhá; jednou naskenovaný dokument tak není nutné znovu digitalizovat. U každé publikace je z těchto důvodů v metadatech a následně v uživatelském rozhraní uvedeno místo uložení[31].

Mohou nastat také případy, kdy mají knihovny ve svých depozitářích různá vydání téhož díla, takže se v digitálních knihovnách objevuje publikace se stejným názvem, ale odlišnými daty (digitálními obrázky) i metadaty, včetně jedinečného identifikátoru (PID). K těmto případům může při velkých objemech digitalizace docházet, neboť je kontrola dokumentů založena na údajích, které identifikují vydání, nikoli dílo jako takové (identifikace díla může být z různých důvodů nejednoznačná). Pro regulaci procesů digitalizace se používá systém Registr digitalizace[32], který umožňuje zveřejnit záměr digitalizace a údaje o jejím postupu. Systém slouží k automatické i manuální kontrole duplicit, které lze odhalit již ve fázi záměru digitalizace.

[30] KNAV, MZK a NK v současné době sdílejí 90 shodných dokumentů, viz např.
https://kramerius.lib.cas.cz/uuid/uuid:18dc4c30-1f67-11e3-a5bb-005056827e52;
https://dnnt.mzk.cz/uuid/uuid:18dc4c30-1f67-11e3-a5bb-005056827e52;
https://ndk.cz/uuid/uuid:18dc4c30-1f67-11e3-a5bb-005056827e52.
V digitálních repozitářích MZK a NK existuje téměř 220 tisíc shodných záznamů.
[31] Viz např. https://ndk.cz/uuid/uuid:ae6ef6fc-435d-11dd-b505-00145e5790ea.
[32] Viz https://www.registrdigitalizace.cz

3.1.1 Digitalizace knihovních fondů v ČR

Definice metadatových formátů (Standardy digitalizace, 2018) slouží jako předpis pro výsledek procesu digitalizace v digitalizačních projektech v Česku a zároveň definují jednotný formát pro paměťové instituce, které chtějí svá data dlouhodobě archivovat v úložišti Národní knihovny ČR. V současné době se do systému pro dlouhodobou ochranu digitálních dat (Long-term Preservation systém, dále též LTP) Národní knihovny ČR přijímají data z digitalizace v NK a MZK a dále data, která byla vytvořena v rámci dotačního programu Veřejné informační služby knihoven (VISK), konkrétně podprogramu VISK 7, který je zaměřen na digitalizaci novodobých textových a zvukových dokumentů, jež jsou uloženy ve fondech knihoven a paměťových institucí a mohou být ohroženy degradací papíru nebo fyzického nosiče. Data přijímaná v rámci programu VISK 7 do systému LTP NK ČR se zároveň zpřístupňují v Národní digitální knihovně (bez data).[33]

3.2 Data, metadata a paradata

Ve spojitosti s digitálními knihovnami a výzkumem pomocí metod digitálních humanitních věd vzniká otázka, s jakými typy dat se uživatelé setkávají. Pro potřeby projektu DL4DH se v rámci této metodiky rozlišují tři typy údajů, které lze charakterizovat následovně:

- data – data v informatice (např. číslo, text, obrázek, zvuk) jsou údaje zaznamenané v digitální (číselné) podobě určené k počítačovému zpracování[34]
- metadata – „…strukturovaná data, která nesou informace o primárních datech … “[35]
- paradata – „doprovodná metadata s informacemi o metodách a technikách získání dat“[36]

Konkrétně se v případě digitálních knihoven za data pokládají obrazy původních publikací, popř. dokumenty PDF, a z nich odvozené texty (ve formátu ALTO, popř. prostého textu), metadata představují údaje o digitalizovaných dokumentech (např. bibliografické záznamy). Paradata popisují nástroje (název, verze) a procesy (parametry, čas spuštění, délka zpracování apod.), s jejichž pomocí data, případně metadata vznikla.

[33] V brzké době bude součástí systému LTP NK ČR i ukládání a publikace born-digital dokumentů.
[34] Srov. Data (informatika), 2001.
[35] Viz Metadata, 2011.
[36] Viz Kučerová, 2019; původně se definice týkala sociologických výzkumů. V této metodice se jako paradata označují jakékoli údaje, které identifikují prostředky a procesy, s jejichž pomocí data vznikla. Jedná se např. o softwarovou aplikaci, která provedla rozpoznání textu; nastavené parametry procesu rozpoznání; čas, kdy k rozpoznání došlo apod.

Vymezení těchto pojmů závisí i na předmětu, který je cílem popisu, analýzy nebo interpretace. Proto např. na dokumenty ve formátu ALTO můžeme nahlížet rovněž jako na kombinaci dat (textu díla) a metadat (údajů o struktuře textu, jeho umístění na straně, řezu písma apod.).

3.3 Digitální knihovna Kramerius

Kramerius je softwarové opensourcové řešení pro zpřístupnění digitálních dokumentů, především digitalizovaných knihovních sbírek. I když převažují periodika a monografie, využívá se také pro zpřístupnění starých tisků, map, hudebnin, grafik, zvukových a dalších typů dokumentů včetně tzv. born-digital publikací. Veškeré dokumenty jsou obohaceny o popisná metadata na úrovni titulu (odpovídající záznamu v knihovním katalogu), jednotlivých stran i mezilehlých úrovní (ročník nebo číslo periodika apod.). Ke stranám jsou u monografií a periodik k dispozici i textové přepisy, s výjimkou starých tisků, kde nástroje pro rozpoznání textu neposkytovaly kvalitní výsledky, a to zejména u českých tisků (pro latinu a němčinu přepisy často existují). V budoucnu lze očekávat doplnění textových přepisů díky využití systému PERO OCR. U některých dokumentů jsou dostupná i metadata popisující logické části dokumentů, jako jsou kapitoly nebo články.

Systém Kramerius sestává z jádra systému a webového klienta, který s jádrem komunikuje pomocí programového rozhraní REST API. Webový klient zajišťuje uživatelské rozhraní pro vyhledávání v metadatech i v plných textech dokumentů a slouží také pro kontinuální čtení publikací. K dokumentům a jejich metadatům je ale možné přistupovat i přímo prostřednictvím zmíněného REST API, které je zdokumentované v repozitáři projektu na platformě GitHub.[37]

Přístup veřejnosti k obrazovým a textovým datům závisí na aktuálně platné autorskoprávní legislativě (viz kapitola 2.5). V současné době pracuje systém Kramerius tří největších digitálních knihoven (NK, MZK, KNAV) se třemi způsoby zpřístupňování svého obsahu (srov. Richter, 2020):

- Veřejné dokumenty – dokumenty kompletně dostupné bez omezení včetně všech metadat. Jedná se buď o díla, která jsou z hlediska autorského práva tzv. volná, nebo taková díla, u kterých má instituce provozující Krameria uzavřené s vlastníkem práv smlouvy o zpřístupnění dat.

[37] https://github.com/ceskaexpedice/kramerius/wiki/

- Díla nedostupná na trhu (DNNT) – díla zařazená na seznam děl nedostupných na trhu. Lze je prohlížet pouze po přihlášení vzdáleně (pro licence typu DNNTO), nebo prostřednictvím terminálu v autorizované knihovně (pro licence typu DNNTT). Obrazová data jsou dostupná jen ve webovém klientu, nelze k nim přistupovat prostřednictvím API voláním odpovídajících koncových bodů. Plné texty dostupné nejsou a z metadat jsou dostupná jen metadata popisná.
- Neveřejné dokumenty – dokumenty, které lze v souladu s autorským zákonem prohlížet jen v budově instituce. Dostupná jsou pouze popisná metadata a náhledy stran.

Digitální knihovny provozované v systému Kramerius obsahují následující data a metadata pro jednotlivé publikace, jejich části (strany), případně vyšší celky (časopisy apod.).

3.3.1 Obrazová data

Obrazová data jsou uložena obvykle ve formátu JPEG 2000 a zpřístupňována ve formátu JPEG. V závislosti na čase a způsobu digitalizace v jednotlivých institucích se objevují různá rozlišení a rozmanité barevné škály. Vícestránková publikace je dostupná ke stažení ve formátu PDF. Na základě nastavení systému Kramerius lze uložit kompletní publikaci, nebo ji postupně stáhnout po jednotlivých sadách stran. Toto omezení zabraňuje přetěžování systému. Born-digital dokumenty jsou k dispozici v formátu originálních PDF a na jejich stahování se vztahují stejné podmínky jako na ostatní druhy digitálních dokumentů.

3.3.2 Textová data

Textová data jsou k jednotlivým stranám dostupná ve dvou základních formátech, buď jako prostý text, nebo jako strukturovaný soubor XML ve formátu ALTO[38]. ALTO je součástí materiálů k dokumentům postupně od přelomu let 2011 a 2012, dokumenty digitalizované do této doby mohou obsahovat textová data pouze v podobě prostého textu. S výjimkou born-digital dokumentů jsou textová data výsledkem automatického rozpoznávání znaků, a to pomocí softwaru, který byl na pracovišti během digitalizace k dispozici. Obvykle jde o různé verze ABBYY Recognition Serveru[39] (a jeho předchůdce, popř. následovníky) nebo nověji o systém PERO OCR[40].

[38] https://github.com/altoxml
[39] https://www.abbyy.com/company/news/announcing-abbyy-recognition-server/
[40] https://github.com/DCGM/pero-ocr

3.3.3 Popisná a další metadata

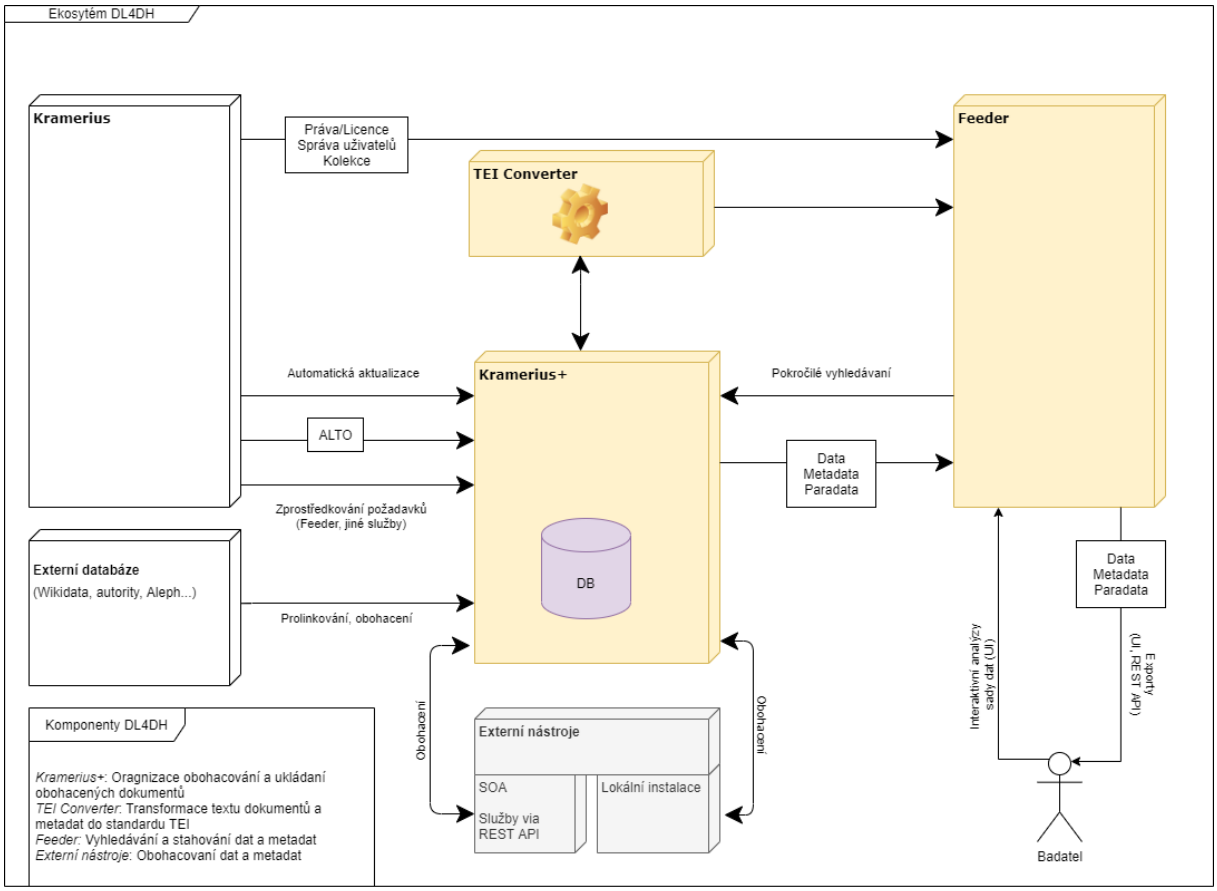
Každý dokument v Krameriu je doplněn o popisná metadata ve schématech MODS[41] a Dublin Core[42]. Tato metadata vznikají podle definic metadatových formátů Národní digitální knihovny (Standard NDK[43], viz kapitola 3.1.1) a pravidel pro popis dokumentů[44]. Metadata ve formátu Dublin Core jsou vždy odvozena z MODS a jsou vždy informačně chudší.

Kromě popisných metadat je uživatelům k dispozici také přehled polí, se kterými pracuje index Apache Solr[45] a popis struktury a rozložení digitalizovaného dokumentu ve formátu JSON-LD dle specifikace IIIF Presentation API v3[46] díky využití brány která pro vytvoření manifestu IIIF Presentation využívá API Krameria Dostupné je také FOXML (Fedora Object XML)[47], které slouží jako interní formát systému Kramerius. Ve FOXML jsou zahrnuta jak popisná metadata, tak metadata administrativní a strukturální. FOXML jednotlivých stran pak obsahuje i datové proudy všech součástí digitalizovaného objektu, tj. např. prostý text nebo zachycení struktury stránky standardem ALTO, pokud nejsou tato data uložena odděleně – v takovém případě na ně jen odkazuje.

3.4 Řešení DL4DH

Řešení, které vzniklo v rámci projektu DL4DH, reaguje na požadavky badatelů v oblasti digitálních humanitních věd. Přihlášení uživatelé mají přístup k velkým objemům dat v textové a obrazové podobě, včetně metadat k těmto objektům. Vše je přístupné jak prostřednictvím uživatelského rozhraní, tak pomocí programového rozhraní REST API[48]. Základní prvky a jejich interakci v rámci implementovaného systému přibližuje následující schéma:

[41] https://www.loc.gov/standards/mods/
[42] https://www.dublincore.org
[43] https://standardy.ndk.cz/ndk/standardy-digitalizace
[44] https://standardy.ndk.cz/ndk/standardy-digitalizace/metadata
[45] https://solr.apache.org
[46] https://iiif.io/api/presentation/3.0/
[47] https://wiki.lyrasis.org/pages/viewpage.action?pageId=66585857
[48] OpenAPI se automaticky generuje z tříd programového kódu v Javě, viz https://bit.ly/dl4dh-api-kramerius-plus a https://bit.ly/dl4dh-api-feeder.



Obrázek č. 1 Základní prvky řešení DL4DH

Uživatelé mohou pomocí dotazů na bibliografické údaje, textový obsah nebo další metadata vyhledat dokumenty, které jsou pro jejich výzkum relevantní. Takto formulované dotazy je možné uložit a sdílet nebo použít později, např. pokud se rozroste prohledávaná sada digitalizovaných publikací nebo je potřeba parametry dotazu upravit. Tyto funkce má na starosti modul DL4DH Feeder.

Textová data se během přípravy pro další zpracování pomocí nástrojů digitálních humanitních věd postupně obohacují s využitím externích, popř. interních služeb a nástrojů. Zejména se jedná o rozdělení do vět, lemmatizaci[49] a morfologickou anotaci[50], kdy se každému tvaru v textu přiřadí odpovídající reprezentativní podoba (lemma, např. 1. pád jednotného čísla nebo infinitiv) a morfologické údaje (např. pád a číslo daného tvaru). K tomuto účelu slouží aplikace UDPipe 2[51]. Vedle toho se v textu rozpoznávají pojmenované entity (tj. osoby, místa, instituce,

[49] Srov. https://www.czechency.org/slovnik/LEMMATIZACE a https://wiki.korpus.cz/doku.php/:pojmy:lemma.
[50] Srov. https://www.czechency.org/slovnik/ANOTACE.
[51] https://lindat.mff.cuni.cz/services/udpipe/

časové údaje apod.), a to pomocí aplikace NameTag 2[52]. Práci s daty a jejich obohacování se věnuje modul Kramerius plus (dále též Kramerius+).

Takto obohacená data je možné stáhnout v jednom ze standardních textových formátů: CSV[53] (text oddělený čárkou), TSV[54] (text oddělený tabulátorem), JSON[55] (JavaScript Object Notation) a TEI[56] (Text Encoding Initiative). Přípravu a generování výstupu ve formátu TEI obstarává interní modul TEI Converter.

3.4.1 Vztah prostředí DL4DH a Kramerius

Cílem projektu bylo v maximální možné míře využít data, která jsou dostupná prostřednictvím systému Kramerius, tj. zejména obrazové materiály, textová data, strukturní a popisná metadata. Pro ukládání metadat a obohacených textů, které nelze získat ze systému Kramerius, slouží pomocný systém Kramerius+. Existující údaje se tak zbytečně neduplikují v pomocné databázi, pokud pro to neexistují jiné závažné důvody (zejména efektivita a rychlost zpracování dotazů). Systém je navržen tak, že data v systému Kramerius+ mohou pocházet vždy jen z jednoho systému Kramerius.

[52] https://lindat.mff.cuni.cz/services/nametag/
[53] https://datatracker.ietf.org/doc/html/rfc4180
[54] Srov. https://www.iana.org/assignments/media-types/text/tab-separated-values a https://www.loc.gov/preservation/digital/formats/fdd/fdd000533.shtml
[55] https://www.json.org/json-en.html
[56] https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html

4 Předzpracování dat a metadat

Zpracování dat představuje kruciální problém v základním i aplikovaném výzkumu, neboť má zásadní dopad na to, jakým způsobem je možné data analyzovat a interpretovat. Cílem projektu DL4DH bylo, aby data mohlo využívat co nejvíce badatelů s různými zájmy a badatelskými přístupy.

Součástí metadat, která jsou dostupná ve výstupech z DL4DH Feederu, se staly údaje, které splňují požadavky kladené na výsledek procesu digitalizace v digitalizačních projektech v Česku, ale nejsou k dispozici prostřednictvím systému Kramerius. Jedná se především o paradata o procesu zpracování, technických parametrech vstupních dat i samotné digitalizace.

Obsah digitalizovaných děl, tj. texty publikací, byl opatřen lemmatizací, morfologickou anotací a identifikací pojmenovaných entit. Jedná se o výsledky analýz pomocí nástrojů počítačového zpracování přirozeného jazyka (NLP), které mohou sloužit jako vhodný základ pro jakékoliv další zpracování materiálu pro konečnou analýzu. Uživatel může při exportu zvolit, které z obohacujících údajů se do výsledné datové sady dostanou.

Architektura projektu DL4DH samozřejmě umožňuje přidávat další externí aplikace, popřípadě nahradit existující aplikace jinými. Implementované řešení momentálně využívá pro obohacení textových dat níže uvedené nástroje, které se staly součástí aktuální implementace vzhledem k jejich výrazné podpoře českého jazyka.

4.1 UDPipe 2

UDPipe[57] je nástroj vyvinutý a provozovaný jako webová služba velkou výzkumnou infrastrukturou LINDAT/CLARIAH-CZ. Texty tokenizuje, lematizuje a opatřuje morfologickou a syntaktickou anotací. UDPipe 2 je dostupný primárně jako webová služba, protože jde o hlubokou neuronovou síť, která ke svému běhu potřebuje prostředí TensorFlow a pro zpracování rozsáhlejších textů dostatečnou rychlostí má také dosti velké a specifické hardwarové nároky. Pro instalaci přímo u uživatelů je služba k dispozici ve formě kontejneru pro Docker. Software je k dispozici pod licencí Mozilla Public License 2.0[58], datové modely pro jednotlivé jazyky jsou

[57] Web: https://lindat.mff.cuni.cz/services/udpipe/info.php, zdrojový kód: https://github.com/ufal/udpipe.
[58] https://www.mozilla.org/en-US/MPL/2.0/

dostupné pod licencí CC BY-NC-SA[59]. Datové modely[60] existují pro více než šedesát jazyků.

4.2 NameTag 2

NameTag 2[61] je nástroj vyvinutý a provozovaný velkou výzkumnou infrastrukturou LINDAT/CLARIAH-CZ, který dokáže v textu rozpoznat a označit pojmenované entity. Je možné ho spustit jako webový server nebo využívat jako webovou službu. Samotný program je k dispozici pod licencí Mozilla Public License 2.0[62], datové modely pro jednotlivé jazyky jsou dostupné pod licencí CC BY-NC-SA[63]. Datové modely[64] jsou v současné době dostupné pro čtyři evropské jazyky. Entity jsou rozčleněny do několika kategorií a podkategorií, jako jsou osoby, názvy institucí (školy, firmy, státní a mezinárodní instituce…), geografické názvy atp.

4.3 Další nástroje pro extrakci dat, jejich obohacení nebo analýzu

Během přípravných prací na projektu DL4DH byly analyzovány i další nástroje, které je možné při segmentaci a další analýze textu využívat. Ty se nestaly součástí procesu obohacování v projektu DL4DH z různých důvodu: licenční podmínky, složitost nebo časová náročnost implementace, vysoká specifičnost úkolu, který nástroj řeší aj. Přesto považujeme za přínosné uvedené aplikace zmínit a ponechat na badatelích, jestli je při zpracování nebo analýze svých dat využijí. Nástroje jsou rozčleněny do několika kategorií podle materie, s níž pracují. Seznam si nečiní nárok na kompletnost, těmto účelům slouží specializované webové stránky, např. TAPoR 3[65].

4.3.1 Zpracování digitálního obrazu

4.3.1.1 Image Comparator (IMCOMP)

Volně dostupný software Image Comparator[66] je webová aplikace pro automatické porovnání dvojice obrázků (knižních ilustrací) pomocí geometrických a fotometrických transformací. Uživateli nabízí několik možností vizualizace rozdílů, které je možné stáhnout ve formě obrázku. Aplikace využívá volně dostupnou knihovnu Traherne Digital Collator[67], která slouží

[59] https://creativecommons.org/licenses/by-nc-sa/4.0/
[60] https://ufal.mff.cuni.cz/udpipe/2/models
[61] Web: https://lindat.mff.cuni.cz/services/nametag/info.php, zdrojový kód: https://github.com/ufal/nametag.
[62] https://www.mozilla.org/en-US/MPL/2.0/
[63] https://creativecommons.org/licenses/by-nc-sa/4.0/
[64] https://ufal.mff.cuni.cz/nametag/2/models
[65] http://tapor-test.artsrn.ualberta.ca
[66] Web: https://zeus.robots.ox.ac.uk/imcomp/index.html, zdrojový kód: https://gitlab.com/vgg/imcomp.
[67] https://www.robots.ox.ac.uk/~vgg/software/traherne/

k odhalení drobných typografických či textových rozdílů mezi dvěma výtisky téhož vydání, které např. vznikly v důsledku oprav během tiskového procesu v éře ručního knihtisku.

4.3.1.2 VGG Image Search Engine (VISE)

VISE je volně dostupný softwarový nástroj[68] pro vědeckou práci s knižní ilustrací. Aplikace využívá technologii počítačového vidění. Slouží k prohledávání rozsáhlého souboru digitalizovaných dokumentů na základě uživatelem definovaného obrazového výřezu. Uživateli nabízí vyhledávání podle obrazových parametrů, řazení výsledku dotazu podle míry podobnosti a vizualizaci identifikovaných odlišností.

4.3.2 Rozpoznání textu (OCR, HTR)

4.3.2.1 PERO OCR

Projekt PERO[69] využívá nejnovější poznatky počítačového vidění, strojového učení a zpracování jazyka pro převod skenů tištěného nebo psaného písma do textové podoby. Uživatelům nabízí jednak webovou aplikaci pro převod a manuální korekce textu, jednak knihovnu v jazyce Python, která využívá REST API[70] systému pro automatickou extrakci textu bez možnosti manuální korekce. Systém dokáže extrahovat text i ze špatně čitelných nebo ručně psaných dokumentů. Aplikace podává dobré výsledky při rozpoznávání moderního písma, ale i fraktury a kurentu. PERO OCR je optimalizované zejména pro texty v němčině a češtině, včetně starých tisků. Výstupy jsou dostupné ve formátu ALTO, PAGE XML[71] a prostého textu. Systém je možné provozovat i na vlastní infrastruktuře, doporučené je ale využívat jej v podobě online služby.

4.3.2.2 Tesseract

Tesseract[72] je opensourcový nástroj pro optické rozpoznání znaků (OCR) a extrahování textového obsahu z obrazových digitálních dokumentů. Podporuje přes 100 jazyků, Výstupy jsou dostupné ve formátu ALTO, hOCR (HTML), PDF, TSV a prostého textu. Výstupy z Tesseractu mají velmi nízkou kvalitu, pokud vstupní obrázek není upraven v rámci pokročilého předzpracování nebo dostatečně kvalitní (např. výška textu, tzv. x-height, má méně než 20 pixelů, obraz

[68] Web: https://www.robots.ox.ac.uk/~vgg/software/vise/, zdrojový kód: https://gitlab.com/vgg/vise.
[69] Web: https://pero-ocr.fit.vutbr.cz, zdrojový kód: https://github.com/DCGM/pero-ocr.
[70] https://app.swaggerhub.com/apis-docs/LachubCz/PERO-API/1.0.4
[71] https://github.com/PRImA-Research-Lab/PAGE-XML
[72] Web: https://tesseract-ocr.github.io, zdrojový kód: https://github.com/tesseract-ocr/tesseract.

je otočený nebo sešikmený, obsahuje tmavé okraje, nejsou odfiltrované nízkofrekvenční změny jasu apod.). Volně dostupnou online implementaci Tesseractu nabízí projekt INDIHU.[73]

4.3.3 Identifikace částí textu

4.3.3.1 Grobid

Aplikace Grobid[74] v jazyce Java umožňuje transformovat nestrukturovaný text ve formátu PDF do strukturované podoby. Extrahuje z publikací vědecké jednotky, jako jsou například bibliografické odkazy, autoři nebo abstrakt, a to pomocí strojového učení.

4.3.3.2 Tabula

Tento nástroj[75] slouží k extrakci tabulek z textových dat ve formátu PDF prostřednictvím jednoduchého webového rozhraní. Uživatel nahraje soubor PDF, manuálně označí hranice tabulky na stránkách a Tabula se pokusí označené tabulky extrahovat ve formátu CSV.

4.3.4 Zpracování přirozeného jazyka

4.3.4.1 GeoNames

Webová aplikace[76] poskytuje veřejně přístupné REST API[77] pro hledání lokalit na základě textového vstupu. Výsledkem hledání je množina míst seřazená podle relevance. Počet dotazů prostřednictvím webové služby je omezen na 20 000 denně. Databáze získává údaje od dobrovolníků, takže data mohou být nepřesná. Kromě geografických údajů o geografické šířce a délce obsahují také příslušnost k vyšším územním celkům, počet obyvatel apod. Názvy lokalit se objevují v národních jazycích. Data je možné stáhnout a využívat off-line pod licencí CC BY[78].

4.3.4.2 Stanza

Volně dostupný balíček skriptů v jazyce Python[79] pro zpracování textů přirozeného jazyka. K dispozici jsou datové modely pro morfologickou analýzu a lemmatizaci 68 jazyků. Rozpoznávat entity je možné v 16 jazycích. Software je dostupný pod licencí Apache Licence, Version 2.0,[80] datové modely mají vlastní licence, obvykle Creative Commons.

---

[73] https://ocr.indihu.cz
[74] Web: https://grobid.readthedocs.io/en/latest/, zdrojový kód: https://github.com/kermitt2/grobid.
[75] Web: https://tabula.technology, zdrojový kód: https://github.com/tabulapdf/tabula.
[76] https://www.geonames.org
[77] https://www.geonames.org/export/web-services.html
[78] https://creativecommons.org/licenses/by/4.0/
[79] Web: https://stanfordnlp.github.io/stanza/, zdrojový kód: https://github.com/stanfordnlp/stanza.
[80] https://www.apache.org/licenses/LICENSE-2.0

5 Architektura systému DL4DH

Následující kapitola popisuje architekturu implementovaného řešení, přibližuje jednotlivé komponenty a jejich roli v systému. Fungování systému na úrovni komponent znázorňuje schéma v příloze č. 12.1.

5.1 Kramerius plus (Kramerius+)

Modul Kramerius+ slouží k ukládání obohacených textových dat a metadat nacházejících se v systému Kramerius. Jedná se o data a metadata získaná z externích nástrojů, jako jsou UDPipe 2 a NameTag 2, z balíčků NDK a z formátu ALTO. Tyto údaje slouží k lepšímu vyhledávání digitálních dokumentů, filtrování jejich obsahu a k exportu.

Kramerius+ poskytuje programové rozhraní REST API pro obohacování publikací a rozhraní pro přístup k databázi obohacených dat.

5.1.1 Moduly

Systém je rozdělen do několika modulů. Toto řešení zajistí vyšší škálovatelnost, zlepší opakované použití a udržitelnost kódu. Každý modul odpovídá za samostatnou funkční část systému. Řešení sestává ze čtyř částí.

5.1.1.1 krameriusplus-core

Modul implementuje datový model, perzistenci dat, zabezpečuje komunikaci s datovými úložišti.

5.1.1.2 krameriusplus-service

Modul obsahuje implementaci obchodní logiky, komponenty pro plánování úloh a komunikaci s ActiveMQ[81].

5.1.1.3 krameriusplus-api

Modul zprostředkovává programové rozhraní REST API pro využití dostupných funkcí.

5.1.1.4 krameriusplus-app

Modul obsahuje konfigurační soubor a hlavní třídu pro spuštění aplikace.

[81] https://activemq.apache.org

5.1.2 Datový model

Datový model Krameria+ (viz příloha č. 12.2) je implementovaný tak, aby požadavky na hierarchii digitálních objektů byly co nejvolnější. Umožňuje vytvářet libovolné hierarchie digitálních objektů: každý objekt může obsahovat seznam potomků – jiných digitálních objektů, popřípadě seznam stran.

Typ digitálního objektu je uložen ve vlastnosti *model* a určuje se při prvotním stahování dat z Krameria. Není-li model z Krameria definován v systému Kramerius+, proces skončí chybovým stavem; na základě analýzy se určí způsob jeho zpracování a zvolená implementace se doplní do systému.

5.1.3 Datové úložiště

Kramerius+ využívá dvě datová úložiště. Pro ukládání obohacených publikací a jejich metadat se využívá nerelační databáze MongoDB[82]. Pro správu úloh, plány úloh, interní logiku procesů, správu exportů a souborů slouží relační databáze PostgreSQL[83].

5.1.4 Úlohy

Pro zajištění plynulosti procesů, robustnosti a ochranu proti zahlcení systému jsou všechny výpočetně náročnější procesy implementované pomocí úloh. Úlohy se po vytvoření zařadí do fronty a postupně se asynchronně spouštějí, jakmile jsou dostupné výpočetní zdroje.

Úlohy se vytvářejí buď samostatně, nebo jako součást plánu úloh. Po vytvoření se úloha uloží do databáze a odešle se do fronty ActiveMQ. Systém v samostatném vlákně automaticky sklízí a spouští úlohy v této frontě. Počet paralelně zpracovávaných úloh se nastavuje v konfiguraci systému.

Úlohy jsou rozdělené na jednotlivé kroky. Každý následující krok se spustí po úspěšném dokončení kroku předchozího. Dojde-li k chybě, systém uloží chybovou zprávu a stav úlohy nastaví na hodnotu FAILED. Takovou úlohu je možné restartovat. Při vykonávání úlohy se postupně ukládá stav jednotlivých kroků, takže se lze v případě chyby vrátit k poslednímu úspěšnému kroku.

[82] https://www.mongodb.com
[83] https://www.postgresql.org

Úlohy je možné vytvářet v rámci plánu úloh. Jeden plán definuje množinu úloh a pořadí, v jakém se mají vykonat. Skončí-li úloha úspěšně, systém automaticky vytvoří následující úlohu z plánu a zařadí ji do fronty.

5.1.5 Obohacení

Publikace se obohacují voláním externích a interních služeb. Každá služba použitá pro obohacení implementuje odpovídajícího obohacovatele, který zajistí kompletní zpracování dat dané služby. Tito obohacovatelé se používají v jednotlivých úlohách.

5.1.6 Export

Proces exportu je kvůli předpokládanému objemu ukládaných dat rovněž řešen pomocí úloh. Výsledkem úspěšné úlohy je fyzický soubor v souborovém systému. Tento soubor se po definovaném čase může automaticky smazat.

5.2 TEI Converter

Jedná se o samostatně fungující modul, který poskytuje služby prostřednictvím rozhraní REST API (primárně zpřístupněného pouze pro modul Kramerius+), ke kterému bude možné v budoucnu připojit i jiné externí aplikace než modul Kramerius+. Rozhraní API je popsáno pomocí specifikace OpenAPI. Primárním vstupem pro koncové body API je objekt ve formátu JSON přizpůsobený exportu z modulu Kramerius+. Na výstupu je vždy dokument ve formátu XML.

TEI Converter poskytuje služby zajišťující převod obohaceného obsahu, který je uložen v systému Kramerius+, do formátu TEI. Konverze probíhá ve třech krocích:

1) Konverze metadat publikace na hlavičku TEI – hlavička (<teiHeader>) se ukládá v systému Kramerius+ pro objekt publikace.
2) Převod obohaceného obsahu na tělo TEI – tělo (<body>) pro každou stránku publikace se ukládá v objektu pro danou stránku publikace.
3) Sloučení hlavičky a těla jednotlivých stránek do jediného dokumentu TEI. V tomto kroku se odstraňuje nevyžádané obohacení (díky filtrování výsledných dat podle fasety NameTag a UDPipe).

Pro každý krok je vytvořen jeden koncový bod. Výstup z posledního koncového bodu se kontroluje vůči schématu XML ve formátu XSD, které vzniklo na základě dokumentace ODD (dokument, který formálními prostředky popisuje povolené prvky v dokumentu TEI).

5.3 DL4DH Feeder

Tato komponenta je určena pro interakci badatelů se systémem. Slouží zejména ke zpřístupnění uložených dat a metadat, popř. ke správě požadavků na jejich vytváření. Údaje o využívání této součásti se zaznamenávají kvůli následné analýze chyb a vylepšování služeb.

5.3.1 Feeder Frontend

Jedná se o komponentu s grafickým uživatelským rozhraním, která slouží k vyhledávání a zobrazení dokumentů v systému Kramerius a Kramerius+. Feeder Frontend volá rozhraní modulu Feeder Backend.

Grafické rozhraní je určeno výzkumným pracovníkům pro snadný uživatelský přístup k systému. Umožňuje vyhledávání, pokročilé dotazování, fasetové filtrování, zobrazení nalezených publikací a export dat ve formátech určených pro strojové zpracování.

5.3.2 Feeder Backend

5.3.2.1 Scheduler

Scheduler zajišťuje hladký chod systému a zamezuje jeho přetížení složitými uživatelskými dotazy. Tato komponenta řadí náročné dotazy do fronty a zajišťuje, aby byly prováděny v méně vytíženém čase. Vyhodnotí-li se dotaz jako málo náročný, provede se okamžitě.

5.3.2.2 Apache Solr

Apache Solr je hlavní komponentou, která se stará o provádění dotazů. Uživatel díky ní může využívat rozsáhlé a pokročilé parametry dotazování, které Solr nabízí. Komponenta slouží pro vyhledávání, filtrování a seskupování. Data získává prostřednictvím komponenty Indexer (viz oddíl 5.3.2.5).

5.3.2.3 Joiner

Komponenta se stará o spojení příchozích dat z Krameria a obohacených dat z Krameria+. Vzhledem k tomu, že jsou data rozdělena do dvou úložišť, umí je tato komponenta sloučit do jednoho dokumentu.

5.3.2.4 QueryProcessor

QueryProcessor je hlavní komponentou zajišťující správu procesu dotazování. Zpracování se liší podle toho, zda dotaz směřuje na obohacené údaje, nebo na prvotní údaje z Krameria. V prvním případě se dotaz posílá na systém Kramerius a výsledek (seznam publikací) se pomocí komponenty Joiner doplní o obohacující údaje. V druhém případě se dotaz přesouvá na komponentu Scheduler, která pracuje nad komponentou Solr.

5.3.2.5 Indexer

Indexer poskytuje rozhraní pro indexování nového obsahu do Solru. Zajistí čištění a případnou transformaci příchozích dat před jejich odesláním do komponenty SolrJ, která obsah indexuje. Indexer se v pravidelných intervalech dotazuje Krameria+ na publikace, které od poslední synchronizace prošly procesem obohacení a zveřejnění. Uživatel s administrátorskými právy může tuto událost spustit také z uživatelského prostředí.

5.3.2.6 Exporter

Komponenta Exporter rozlišuje, zda se jedná o export jedné strany s kompletním ukládaným obsahem, nebo o ostatní případy (export více stránek nebo ukládání vybraných částí obsahu). V prvním případě si Exporter vyžádá daný formát z Krameria, popř. Krameria+ synchronním voláním a výsledek vrátí uživateli.

Při exportu více stran nebo při použití filtru, který modifikuje podobu výstupu, komponenta požádá asynchronním voláním Kramerius+ o přípravu exportu a uloží identifikátor úlohy.

Fyzicky se exportované soubory ukládají v systému Kramerius+ a komponenta Exporter na ně pouze odkazuje.

Exporter také vrací informace o stavu exportů jednotlivých uživatelů. Tyto stavy mohou nabývat následujících hodnot: *čeká na zpracování, zveřejněný, nedostupný* (došlo ke smazání souboru z úložiště Krameria+)*, chyba při zpracování*.

5.3.2.7 PostgreSQL

PostgreSQL je databázový systém, který Feeder Backend využívá pro údaje o exportech uživatelů, o interních uživatelských účtech a jejich rolích. V databázi se uchovává datum poslední synchronizace se systémem Kramerius+ a interní statistiky o dotazech.

6 Datové sady

Účelem sady nástrojů DL4DH je zpřístupnění velkého souboru dat, které knihovny vytvořily v procesu digitalizace svých sbírek, pro vědecký výzkum. V humanitních vědách stále převažují kvalitativní výzkumné metody založené na interpretativních a hermeneutických principech. Existence digitálních dat však humanitním badatelům umožňuje uplatnit při analýze kulturních a jazykových jevů širší spektrum výzkumných postupů, včetně kvantitativních. Při práci s daty je pak důležité rozlišovat jednotlivé typy dat, které vymezují, jaké konkrétní techniky lze při analýze uplatnit.

6.1 Typy dat

Základní členění spočívá v rozlišení mezi strukturovanými a nestrukturovanými daty. Typickým příkladem nestrukturovaných dat je text v podobě textového řetězce znaků. V digitálních humanitních vědách se tedy s nestrukturovanými daty setkáváme velmi často a v současné době se rozšiřuje množství tzv. „dolovacích“ technik a algoritmů, které na vstupu předpokládají právě souvislý textový řetězec. Pro strukturovaná data naopak existuje datový model, který popisuje a jasně vymezuje každou hodnotu v datovém souboru. Při práci s daty je obvykle transformujeme tak, aby odpovídala datovému modelu, který naše analýza předpokládá. I techniky pro práci s nestrukturovanými daty jsou založené na hledání implicitní struktury. V případě textů tak většina postupů vychází z rozložení souvislých řetězců na jednotlivé tokeny, které můžeme různě vymezit např. dle jejich příslušnosti k textům nebo na základě jejich pozice v řetězci.

Další užitečnou perspektivu poskytuje rozlišení dat, metadat a paradat. Za data se mohou považovat datové objekty (v DL4DH např. stránka knihy), přičemž metadata, resp. paradata obsahují popis každého objektu (kniha, ze které stránka pochází; pozice stránky v knize), resp. proces jejich vzniku (software použitý k rozpoznání textu na stránce apod.). V badatelské praxi je možné se mezi úrovněmi dat a metadat pohybovat. Pokud se za data považují jednotlivé textové dokumenty, které jsou v metadatech popsány např. svou příslušností k vydavatelskému formátu jako periodický tisk, nebo kniha, je možné tuto metadatovou informaci použít k vytvoření binární proměnné „periodikum“, která bude nabírat hodnotu 1 pro dokument, který je periodikem, a 0 v ostatních případech. Bohaté metadatové informace tak výrazně rozšiřují pole dostupných analýz a jsou nezbytné pro supervizované metody strojového učení.

A konečně při práci s daty je třeba mít na paměti rozdíly mezi numerickými a kategorickými proměnnými. Numerická proměnná nabírá jako své hodnoty čísla, která mohou být spojitá, nebo diskrétní (celá čísla). Podílem se může vyjádřit třeba výskyt konkrétního slova v poměru k celkovému počtu slov v dokumentu, zatímco samotný počet slov v dokumentu bude vždy celým číslem. Kategorické proměnné jsou takové, jejichž hodnota zachycuje příslušnost konkrétního pozorování k dané kategorii (např. jazyk dokumentu). Pokud má proměnná jen dvě kategorie, jedná se o binární proměnnou (viz výše). U vícekategoriálních proměnných lze někdy stanovit pořadí kategorií, v takovém případě pak hovoříme o ordinálních proměnných. Typ proměnné zpravidla určuje rozdělení hodnot a vymezuje tak aplikovatelné statistické metody. Např. u kategorických proměnných obvykle nelze spočítat statistiku průměru, byť u ordinálních proměnných je i takový postup v zásadě možný.

Numerické proměnné lze stanovením intervalů převádět na kategorické – číselnou proměnnou „délka textu“ změřenou počtem znaků tak lze převést na kategorickou proměnnou s hodnotami „krátký text“ a „dlouhý text“ tím, že stanovíme hranici počtu znaků, od níž budeme text považovat za dlouhý. Opačným směrem (z kategorických proměnných na numerické) data transformovat nelze přímo, ale protože i u kategorických proměnných lze počítat jejich četnosti výskytu, mohou se tyto četnosti stát numerickými proměnnými, pokud jsou data agregována na vyšší úroveň. V analýze textů jde o běžný postup: na začátku se pracuje na úrovni jednotlivých výskytů tokenů, později se jednotkou, která je vyjádřena jako číselná proměnná, stává místo tokenu typ (slovo) nebo dokument. Numerické proměnné vytvořené z četností kategorických proměnných však zpravidla budou mít nenormální rozložení. Ve standardních výstupech DL4DH převažují vzhledem k povaze dat kategorické proměnné.

6.2 Struktura dat

Data lze uchovávat v různých strukturách. Elementární formou strukturovaných dat je tabulka, kde každý řádek představuje jeden záznam nebo pozorování a každý sloupec definuje typ naměřené hodnoty. Tato mřížková struktura odpovídá schématu numerických matic, s nimiž pracují statistické modely. I proto patří tabulka k nejběžnějším způsobům uspořádání dat. Kromě naměřených hodnot pro různé proměnné je možné se setkat i s daty, která mají podobu časových řad, grafů (sítí) nebo geolokačních koordinát. I tyto typy dat však lze reprezentovat jako tabulku nebo jejich kombinaci. Například síť lze popsat pomocí soustavy dvou tabulek, kdy jedna obsahuje pouze jeden sloupec zaznamenávající vrcholy grafu a druhá ve dvou sloupcích na každém řádku uchovává informaci o každé existující hraně mezi vrcholy. Nástroje DL4DH

poskytují badatelům základní a obohacená data ve strukturované podobě a s metadatovým popisem. Další transformace a agregace dat nebo vytváření nových proměnných na základě existujících se odvíjejí od potřeb konkrétního výzkumu a pracovního postupu.

6.3 Datové formáty

Datové formáty představují konvence používané k reprezentaci dat v elektronické podobě. Pro sdílení dat je nejjednodušší uložit data do souboru, který lze snadno sdílet a dále zpracovávat.

Zvolená reprezentace může mít podobu binárního nebo textového souboru. V prvním případě (např. obrazy ve formátu JPEG) je data nutné při ukládání a po přečtení nějakým způsobem interpretovat, o což se stará specializovaný software. Textové soubory mají podobu sekvence písmen, číslic a speciálních znaků (např. tabulátorů) a může je číst a interpretovat i člověk, pokud rozumí zvolenému uspořádání údajů v textu.

Při ukládání textového souboru je třeba zvolit vhodnou kódovou stránku, která zajistí, že se všechny používané znaky správně uloží a po otevření dokumentu opět načtou. V současné době se při zaznamenávání znaků vychází ze standardu Unicode (2021), který definuje podobu jednotlivých znaků i jejich počítačovou reprezentaci (tj. strojově čitelnou sekvenci nul a jedniček). Pro ukládání souborů s texty přirozeného jazyka je vhodné zvolit dvoubajtové kódování UTF-16 nebo kódování UTF-8, které používá proměnnou délku znaku od 1 do 4 bajtů, a je tedy úspornější.

Soubory s textovými daty a metadaty, které jsou dostupné prostřednictvím nástrojů DL4DH, používají kódování UTF-8 bez signatury. Badatelé mají k dispozici následujících pět formátů.

6.3.1 Prostý text

Prostý text, tj. zachycení jednotlivých znaků kontinuálního textu bez formátovacích informací, představuje jeden z hlavních formátů využívaných při počítačovém zpracování přirozeného jazyka (NLP). V rámci nástrojů DL4DH vzniká prostý text na základě automatického rozpoznávání znaků (OCR), a to extrakcí z formátu ALTO (viz následující kapitola).[84]

6.3.2 ALTO

Digitalizované dokumenty většinou procházejí procesem automatického rozpoznávání znaků (OCR), jehož výsledkem je sada dokumentů ve standardu ALTO (Analyzed Layout and Text

[84] Jednu stranu u takto transformovaných textů obvykle tvoří jeden souvislý odstavec prostého textu, i když je v originální předloze text rozdělen do více odstavců.

Object), který formálními prostředky jazyka XML zachycuje naskenovanou stránku (její rozměry a okraje) a údaje o umístění textu na stránce a jeho formátování (jak na úrovni textových bloků, tak dílčích úseků). Text stránky je pomocí elementů rozdělen na textové bloky (element <TextBlock>), v jejich rámci na řádky (<TextLine>) a nakonec na minimální rozpoznané úseky[85] ohraničené elementem <String>, který v atributech uchovává jednak rozpoznaný text ve strojově čitelné podobě, jednak údaje o formátování textu[86] a o přesném umístění prvku na stránce. Mezi zachycenými součástmi stránky mohou být také netextové prvky, např. ilustrace (<Illustration>) nebo grafické prvky určené pro oddělení jednotlivých bloků textu (<GraphicalElement>).

Standard ALTO se vyvíjí od roku 2004, poslední aktualizace (verze 4.3) pochází z května 2022. Jedná se o jeden ze standardů Knihovny Kongresu, která jej hostí na svých webových stránkách[87].

6.3.3 TSV, CSV

Formáty TSV (tab-separated values) a CSV (comma-separated values) představují způsob, jak v prostém textu uspořádat strukturovaná data do mřížky. Názvy sloupců (pokud existují) a jednotlivé hodnoty jsou zapsány do buněk tabulky, jejichž pozice je definována řádkem, na kterém se nacházejí, a sloupcem, jehož pořadí určuje počet oddělovacích znaků na řádku. Výhoda dat uspořádaných do mřížky spočívá v tom, že v tomto formátu lze data prakticky bezprostředně použít pro modelování a statistickou analýzu.

Souborové přípony TSV a CSV jednoduše označují použitý oddělovací znak. V případě TSV se jedná o tabulátor, v případě CSV o čárku (comma) nebo středník (někdy se pro hodnoty oddělené středníkem používá přípona CSV2). Jelikož se jedná o nebinární formáty, jsou data ve formátech TSV a CSV čitelná nejen strojově, ale i lidsky. To znamená, že např. stačí vytisknout obsah souboru na obrazovku z příkazové řádky, aby uživatel mohl obsah dat přečíst a porozumět mu. Data oddělená oddělovacím znakem patří k nejrozšířenějším formátům pro výměnu dat a jsou podporována prakticky ve všech analytických nástrojích a programovacích jazycích. V principu lze jako oddělovač zvolit libovolný znak, tabulátory, čárky a středníky jsou pouze standardem, nikoliv nezbytným řešením.

[85] Obvykle jde o samostatná slova včetně navazující interpunkce.
[86] Výstupy z aplikace PERO OCR údaje o formátování textu neobsahují.
[87] Viz https://www.loc.gov/standards/alto/.

Pro práci s přirozeným jazykem mohou tyto formáty působit potíže při načítání dat, protože oddělovače (zejména čárky a středníky) se v textu běžně vyskytují. Textové řetězce v tabulce se proto oddělují od ostatního obsahu použitím uvozovek, přičemž uvozovky použité v textu samotném se zdvojí. DL4DH předává uživatelům tabulková data v podobě inspirované územ z počítačové lingvistiky, kdy každý řádek zaznamenává výskyt jednoho textového tokenu. Jedná se tedy o strukturovanou datovou formu. Pokud je potřeba převést data na nestrukturovaný textový řetězec, musí se vybrat příslušný sloupec a nahradit konce řádků mezerou.

Formáty TSV a CSV nejsou efektivní v případě, že obsahují opakující se informace. Pokud např. data zahrnují unikátní identifikátor dokumentu, v němž se tokeny vyskytují, musí se tato informace opakovat na každém řádku, takže pro zachycení údajů je potřeba větší objem dat. Tímto nedostatkem netrpí formáty typu TEI nebo JSON, které mohou hodnoty uchovávat ve složitějších, hierarchicky strukturovaných datových objektech nebo např. polích.

Formáty TSV a CSV nejsou vhodné tam, kde lze ve zdrojových datech předpokládat opakované údaje, což je běžné například v bibliografických záznamech, kde se může vyskytovat několik desítek autorů jednoho svazku apod. Pokud se nemá exportovat sada navzájem provázaných tabulek, podobná SQL databázi, je nutné při exportu opakované údaje nějakým způsobem ošetřit a při dalším zpracování s takovou úpravou počítat.

6.3.4 JSON

Jak napovídá nezkrácené označení tohoto formátu (JavaScript Object Notation), byl původně určen pro programovací jazyk JavaScript a sloužil k zachycení dat, která mohou být organizována v polích nebo agregována v objektech. Jedná se o textový, jazykově nezávislý formát, který se v současné době využívá pro přenos dat. Je čitelný pro člověka a strojově zpracovatelný, přičemž programové knihovny pro práci s tímto formátech existují pro mnoho programovacích jazyků. Formát je standardizován normou ECMA-404 (The JSON data interchange syntax, 2017).

Pole hodnot či objektů se ohraničují hranatými závorkami (`[]`). Hranice objektu se označují složenými závorkami (`{}`), názvy vlastností jsou uzavřeny v rovných dvojitých uvozovkách,

následuje dvojtečka a přiřazená hodnota (objekt nebo pole hodnot či objektů, číslo, text a speciální hodnoty *true, false* a *null*), pouze textová hodnota je uzavřena od rovných dvojitých uvozovek. Hodnoty a objekty se od sebe oddělují čárkou.[88]

Náležitosti dat ve formátu JSON je možné formálně zachytit pomocí standardu JSON Schema[89], který však stále nedospěl k finální verzi návrhu.

6.3.5 TEI

V oblasti digitálního zpracování textů se prosadily také standardy využívající rozšiřitelný značkovací jazyk (Extensible Markup Language neboli XML, srov. XML Technology, 2015). Za obecně přijímaný a poslední dobou často využívaný se považuje standard konsorcia *Text Encoding Initiative* (TEI) v poslední, páté verzi označované jako TEI P5 (TEI: P5 Guidelines, 2022). Tento standard definuje elementy, jejich atributy, hierarchii a sémantiku takovým způsobem, aby bylo možné zachytit co největší spektrum formálně i obsahově různorodých historických pramenů: od prózy a poezie přes drama až po slovníky. Specifikace rovněž definuje široké spektrum prvků pro popis metadat o pramenném textu. Doporučení TEI P5 využívá například Deutsches Textarchiv (2007–2022), textový archiv německých, převážně tištěných textů z období 17.–19. století, a Text Creation Partnership (bez data), věnující se přepisům zejména anglickojazyčných starých tisků z let 1470–1800, uložených v téměř 150 knihovnách.

V projektu DL4DH se široká paleta elementů a atributů ze standardu TEI omezila na takový repertoár prvků, které dokáží v přijatelné míře obecnosti a podrobnosti zachytit digitalizované publikace různého druhu. Pro tyto úpravy se využily prostředky, které tento standard nabízí, tj. modifikace schématu[90] pomocí ODD (One Document Does it All)[91], a to s využitím nástroje Roma[92]. Kvůli použití nástroje NameTag 2 vznikl převodník[93], který různým skupinám a podskupinám rozpoznaných entit přiřadil adekvátní elementy ze standardu TEI.

[88] Viz např. https://www.json.org/json-cz.html.
[89] https://json-schema.org/specification.html
[90] https://www.tei-c.org/release/doc/tei-p5-doc/en/html/USE.html#MD
[91] https://tei-c.org/guidelines/customization/getting-started-with-p5-odds/
[92] https://romabeta.tei-c.org
[93] https://github.com/LIBCAS/DL4DH/issues/19#issuecomment-848394546

Ukázka dat identické pasáže z publikace s identifikátorem `uuid:0c94cf70-188a-11e4-8f64-005056827e52` zpracovaná ve výše popsaných formátech je součástí přílohy (viz kapitoly 12.3–12.8).

6.4 Export dat

Výstupy ze systému Kramerius+ slouží badatelům k samotné vědecké práci, k níž mohou využít nástroje a postupy, které jsou specifické pro jejich obor nebo výzkumné téma. V tomto ohledu by měla být poskytovaná data univerzálně použitelná. Zároveň by měla v duchu principů FAIR splňovat kritéria, která se týkají interoperability. Proto jsou údaje v datech a metadatech zachyceny s využitím obecných standardů. Badatelé mají údaje k dispozici ve formátech TSV, CSV, JSON, TEI, prostý text a ALTO (viz předchozí kapitola). Stejně tak je možné získat i obrazová data (viz kapitola 3.3.1). Získané údaje obsahují také licenční informace, metadata o publikacích nebo paradata o použitých nástrojích, což vše přispívá k reprodukovatelnosti výzkumu.

6.4.1 Uživatelské rozhraní

Export dat z uživatelského rozhraní umožňuje aplikace DL4DH Feeder. Uživatel nejprve zvolí publikace, jejich části nebo naopak celé kolekce, a to jednoduchým označením zaškrtávacího pole ve výpisu výsledků vyhledávání. Po kliknutí na tlačítko „Export“ se zobrazí dialogové okno, kde uživatel vybere preferovaný formát a požadované atributy, které si přeje exportovat.

Badatel může data získat také programově, a to s využitím programového rozhraní REST API modulu DL4DH Feeder (viz kapitola 7.2.2).

6.4.2 Kolekce a datové sady

Soubory dat mohou mít podobu kolekcí a uživatelských datových sad, které podporují principy FAIR pomocí jednoduchého ukládání a sdílení odkazů na tyto prvky (viz kapitola 6.5). Datová sada je vždy definovaná pomocí seznamu identifikátorů UUID publikací nebo jejich hierarchických částí (např. kořenové číslo, ročník, číslo, svazek), které v sobě vždy obsahují podřízené jednotky datového modelu (např. ročník, číslo, svazek, stránky). Podle původu a způsobu organizace souboru dat se rozlišují:

- **kurátorská kolekce** představuje specifický prvek systému Kramerius, kterým se označuje soubor publikací zastřešený unikátním UUID.[94] Za jejím vznikem stojí poměrně komplexní proces: kolekce vznikají v rámci jednoho repozitáře a orientují se na určité tematické fenomény (např. včelařství nebo mineralogie)[95], přičemž vycházejí z předmětové specifikace zahrnutých děl. Tyto soubory opakovaně procházejí předmětovou kontrolou správce digitální knihovny ve spolupráci s odborníky na dané téma, tj. návrhem, posuzováním stavu a obsahu, zařazením/vyřazením děl. Díla, která jsou pro soubor podstatná a ještě nebyla digitalizována, se zařazují do digitalizačních plánů a průběžně digitalizují, aby vznikl co nejúplnější celek pro dané téma. Kurátorské kolekce jsou součástí systému Kramerius a obsahový správce je může přenést do systému Kramerius+., díky čemuž budou dostupné i z DL4DH Feederu. Tento proces je doporučen pro řízené doplňování a scelování dostupných dat.[96]
- **uživatelské datové sady** – v rámci DL4DH se tímto pojmem označuje sada dat, která obsahuje jednotlivé publikace nebo jejich části identifikované pomocí seznamu UUID. Tyto sady mohou vznikat postupně a zpřesňovat se díky aktivitě uživatelů v rozhraní DL4DH Feederu, a to z dostupných dat na základě filtračních mechanismů a následné badatelské analýzy. Jejich sdílení podporuje reprodukovatelnost výzkumu. Dostatečně kvalitní obecně zaměřená uživatelská datová sada se může stát základem kurátorské kolekce.

6.5 Sdílení dat

Otevřenost dat a jejich digitální dostupnost prostřednictvím on-line nástrojů jsou cestou k implementaci širších konceptů otevřené vědy (open science; srov. Bueno de la Fuente, bez data). Primárním účelem otevřenosti ve vědě je zajistit vyšší efektivitu výzkumu a zvýšit jeho důvěryhodnost. K dosažení těchto cílů vede cesta přes reprodukovatelnost výsledků a transparentnost výzkumného procesu. Nástroje DL4DH byly sestaveny tak, aby aplikaci konceptů otevřené vědy podporovaly. K tomu slouží několik souvisejících funkcí:

[94] V prostředí systému Kramerius se tyto kolekce označují jako virtuální sbírky, resp. sbírky (v seznamu faset).
[95] Viz např. https://kramerius5.nkp.cz/search?collections=vc:285741e3-886f-49fc-8f16-9c0af6e6901d.
[96] Kurátorské kolekce procházejí zásadní proměnou v Krameriu verze 7. Nově jsou koncipovány jako další druh dokumentu, který má vlastní popis ve formátu MODS a jsou specifické tím, že mohou obsahovat libovolné další druhy dokumentů, včetně jiných kolekcí. Mohou tak vzniknout bohatě strukturované kolekce kopírující existující fyzické sbírky nebo kolekce, jejichž struktura přináší novou informační hodnotu.

- **Ukládání a sdílení uživatelských dotazů.** Dotaz použitý pro výběr dat pomocí DL4DH Feederu lze přenášet, předávat dalším uživatelům a pokládat opakovaně vůči stejné sadě digitálních dokumentů, případně se dá totožný dotaz použít pro výběr dat z odlišné sady jako srovnávacího vzorku (např. z jiné instance digitální knihovny Kramerius). Při opakovaném užití dotazů je třeba mít na paměti, že data v DL4DH Feederu se průběžně doplňují a optimalizují, výsledky stejného dotazu použité s určitým časovým odstupem se proto mohou částečně lišit s ohledem na použitou verzi DL4DH Feederu a repertoár dokumentů spravovaných digitální knihovnou. Při využití výsledků je proto třeba uvádět informaci o použitém dotazu, datu jeho použití a o instanci i verzi DL4DH Feederu.
- **Ukládání a sdílení odkazů na datové sady.** Kromě dotazů dovoluje DL4DH Feeder vybírat data rovněž zadáním seznamu identifikátorů UUID publikací nebo jejich částí, které se mají zobrazit a dále zpracovat. Pokud tedy badatel doplní výstupy konkrétního dotazu také seznamem použitých dat ve formě výčtu UUID, může jiný uživatel tento seznam použít pro výběr totožného datového souboru. Seznam UUID je možné exportovat přímo z DL4DH Feederu. I zde platí, že existuje riziko změny datového obsahu spojeného s konkrétními publikacemi, např. v důsledku redigitalizace či aplikace novější verze obohacení plného textu; je proto vhodné při publikaci a sdílení odkazů na datovou sadu uvádět datum vzniku seznamu UUID a odkazovat na použitou instanci a verzi DL4DH Feederu.
- **Export a sdílení dat.** Data v DL4DH Feederu je možné stáhnout a uložit pomocí exportů v různých formátech (viz výše v této kapitole). Sdílení těchto exportů je však omezeno autorskými právy a licencí, na základě které byla data z DL4DH Feederu, resp. z digitální knihovny získána. Export vždy obsahuje metadata s informacemi o této licenci a ustanovení licence je třeba v plném rozsahu dodržovat. Jednotlivé typy licencí mohou omezovat volné šíření dat mimo užití pro potřeby vlastního výzkumu, jiné licence mohou být restriktivní ve vztahu ke komerčnímu užití, některá data však mohou být zcela otevřená a volně šiřitelná. Při jakémkoli sdílení exportovaných dat je nutné tyto licenční údaje zachovat a předávat společně s daty.
- **Dostupnost paradat.** DL4DH Feeder neposkytuje pouze obohacené plné texty a obrazová data. Exportované soubory obsahují také paradata s informacemi o způsobu vzniku dat i metadat. K těmto údajům patří např. název použitého softwaru včetně jeho verze, časový údaj, kdy manipulaci s daty došlo apod. Tyto údaje umožňují pochopit procesy, které vedly k výsledné podobě datového souboru, zohlednit známé nedostatky konkrétních verzí pou-

žitých technologií, případně srovnávat výsledky získané z dat při užití různých postupů jejich tvorby (digitální knihovny obvykle obsahují dokumenty a metadata vzniklé v různých obdobích digitalizace, je proto možné sledovat dopady technologických a procesních změn i v rámci jediné instance DL4DH Feederu).

Všechny tyto funkce a možnosti poskytované nástroji DL4DH jsou podstatné z hlediska praktické implementace postupů otevřené vědy. Souvisí též s opakovanou využitelností, která je jedním ze čtyř základních stavebních kamenů principů FAIR a bývá na rozdíl od vyhledatelnosti, přístupnosti a interoperability v praxi opomíjena.

7 Práce s nástroji DL4DH

Nástroje, které byly vytvořeny pro badatele z oblasti humanitních věd, aby mohli lépe využít digitální knihovny, sestávají ze dvou částí určených odlišným uživatelům. Modul Kramerius plus (Kramerius+) slouží kurátorům digitálních sbírek ke správě procesů, které zajistí obohacení a export dat z digitalizovaných publikací pro potřeby badatelů. Modul DL4DH Feeder slouží badatelům, aby v digitálních sbírkách našli relevantní publikace a získali jejich data v požadovaném formátu pro další výzkum. Oba moduly lze ovládat jednak pomocí uživatelského rozhraní ve webovém prohlížeči, jednak prostřednictvím programového rozhraní REST API. Následující kapitoly popisují hlavní principy práce s těmito moduly, ilustrační snímky obrazovek pocházejí z vývojové fáze systému, takže finální podoba uživatelského prostředí se může lišit. Odkaz na podrobnější nápovědu bude dostupný na webových stránkách jednotlivých modulů, resp. na veřejném úložišti zdrojového kódu.[97]

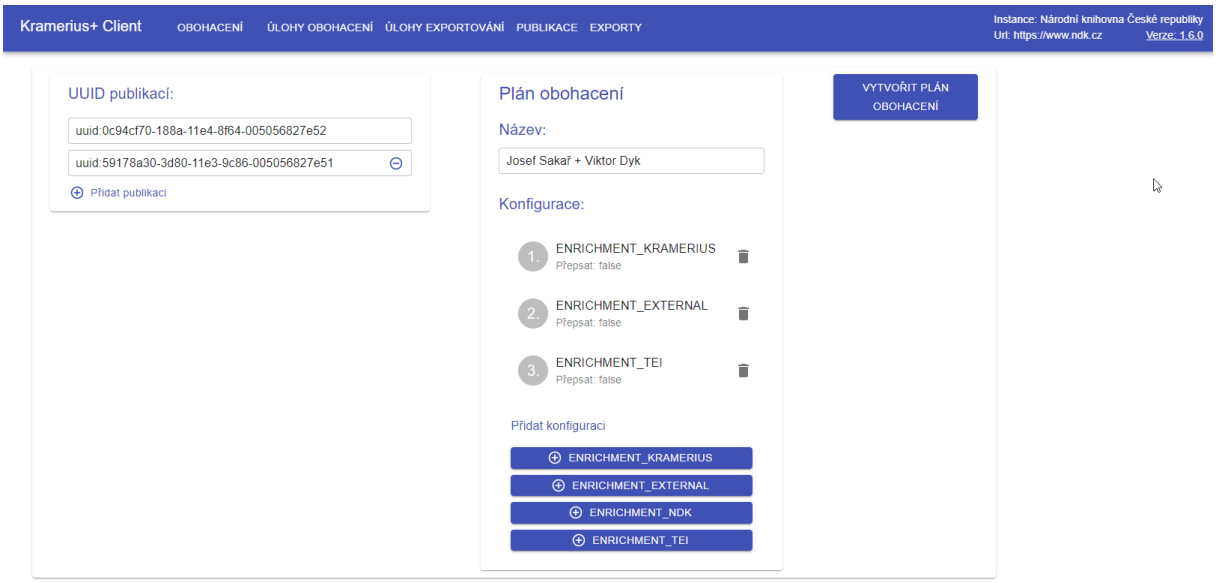
7.1 Kramerius+

7.1.1 Grafické uživatelské rozhraní

Webová aplikace je pomocí horního menu rozdělena na několik částí odpovědných za celý proces obohacování a publikování dokumentů.

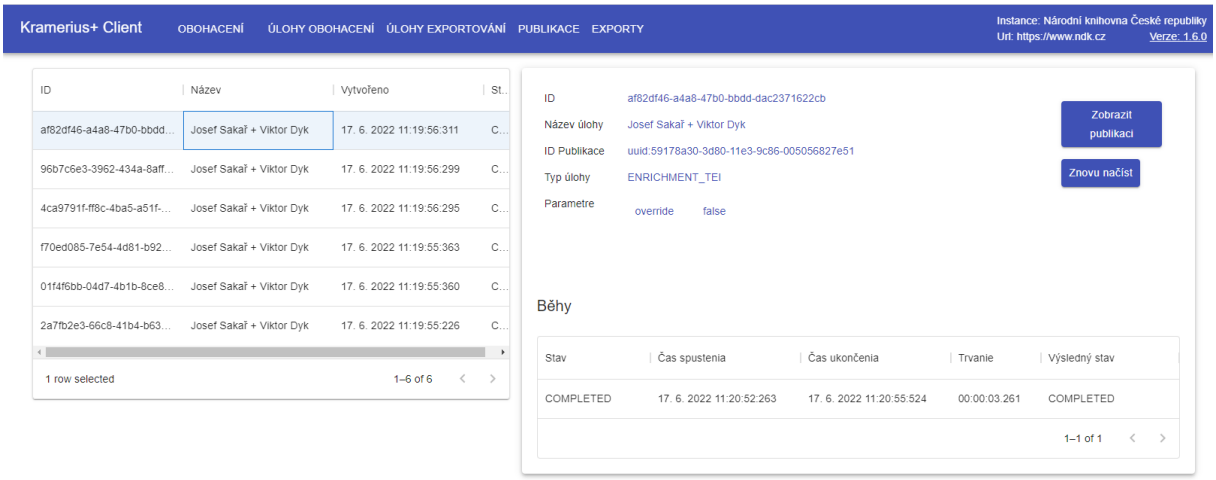
Na stránce *Obohacení* správce nejprve zadá seznam identifikátorů publikací (UUID), které se mají zpracovat, a definuje jednotlivé kroky, jimiž zvolené publikace projdou (získání primárních dat z knihovny Kramerius, obohacení dat pomocí externích služeb, zpracování metadat a paradat z balíčků NDK a konverze dokumentu do formátu TEI). Takto definovaný plán obohacení je vhodné pro další orientaci pojmenovat.

[97] Viz zejména https://github.com/LIBCAS/DL4DH-Kramerius-plus a https://github.com/LIBCAS/DL4DH-Feeder.



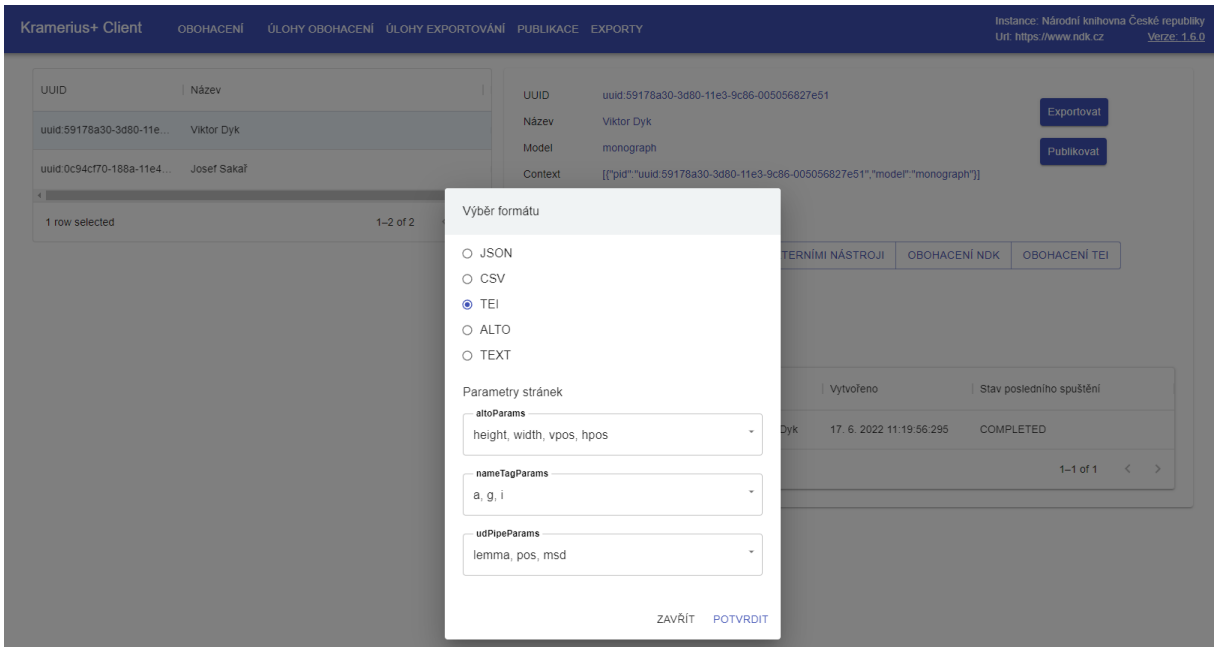
Obrázek č. 2 Ukázka stránky pro obohacení publikací

Na stránce *Úlohy obohacení* jsou vidět jednotlivé kroky, jimiž publikace prošly. Kromě souhrnné informace o ne/úspěšném dokončení jednotlivých kroků jsou k dispozici i podrobnější údaje o dílčích operacích. Krok, který skončil chybou, je z tohoto místa možné opět spustit. Stránka slouží také k přechodu ke zpracovaným publikacím.



Obrázek č. 3 Přehled úloh, jimiž prošla obohacovaná publikace

Na stránce *Publikace* se zobrazuje seznam dokumentů, které prošly zpracováním. Správce může jednotlivá díla exportovat (uložit v různých formátech) a označit jako publikovaná, aby se zobrazila uživatelům v prostředí DL4DH Feederu.



Obrázek č. 4 Nastavení parametrů exportu

Exporty včetně podrobných informací o proběhnuvších dílčích krocích se zobrazují na stránce *Úlohy exportování*. Exportované soubory je možné stáhnout ze stránky *Exporty*.

7.1.2 REST API

Správce se musí nejprve autentifikovat, tj. přihlásit prostřednictvím volání `/api/login`. Pro obohacování publikací jsou k dispozici koncové body na adrese `/api/enrichment/`: jednak `/api/enrichment/plan` pro vytváření pokročilejších scénářů obohacení publikací, jednak samostatné body pro spuštění specifických úloh, např. `/api/enrichment/kramerius`, `/api/enrichment/tei` apod.

Voláním koncových bodů na adrese `/api/jobs/` získá správce přehled o probíhajících i dokončených úlohách, na adrese `/api/jobs/{id}/restart` je možné úlohu, která skončila chybou, znovu spustit.

Koncové body na adrese `/api/publications/` slouží k získání přehledu o publikacích zpracovaných systémem Kramerius+. Kromě seznamu se základními metadaty (`/api/publications/list`) je možné získat údaje o podřízených publikacích (`/api/publications/{id}/children`; např. jednotlivá čísla časopisu) nebo o procesu zpracování jednotlivých stran (`/api/publications/{id}/pages`) a jejich výsledek (`/api/publications/{id}/pages/{pageId}`).

Koncové body na adrese `/api/exports/` správce využije pro získání přehledu o exportovaných publikacích (`/api/exports/list`), pro spuštění procesu exportování s požadovanými parametry (`/api/exports/{id}/csv` apod.) nebo ke stažení souboru, který na základě exportu publikace vznikl (`/api/exports/download/{id}`).

Podrobnější a aktuální informace o programovém rozhraní REST API jsou dostupné v úložišti zdrojového kódu, popř. z URL adresy pro dokumentaci OpenAPI[98].

7.2 DL4DH Feeder

7.2.1 Grafické uživatelské rozhraní

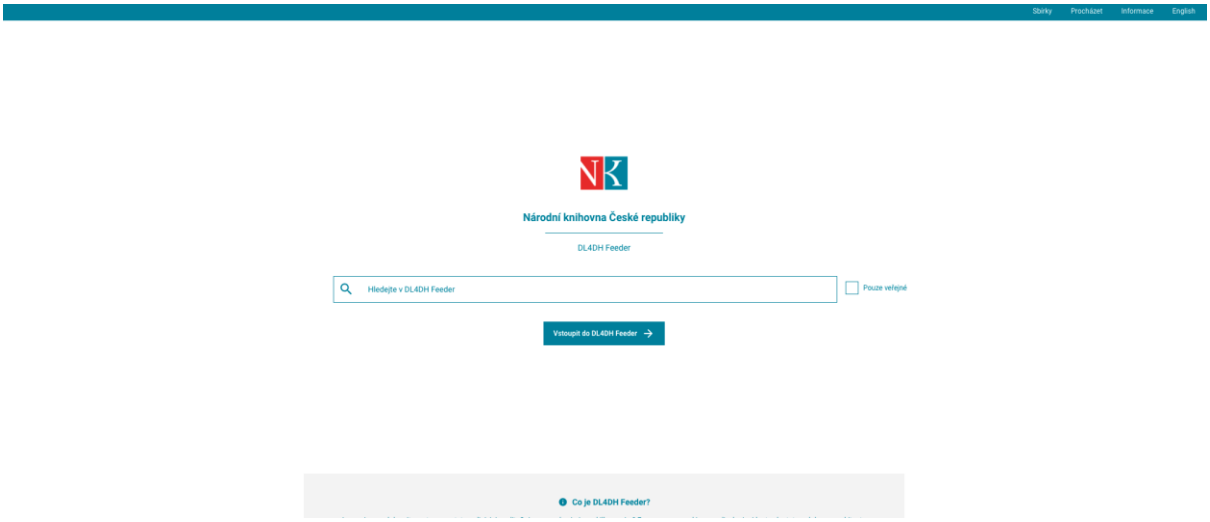
Hlavní komponentu umožňující vyhledávání představuje modul DL4DH Feeder (viz kapitola 5.3), který nabízí výzkumným pracovníkům přehledné grafické rozhraní pro práci s daty uloženými v modulu Kramerius+ i digitální knihovně Kramerius. Vizuální podoba Feederu vychází z webového klienta systému Kramerius určeného běžným uživatelům k vyhledání i prohlížení digitálních dokumentů[99]. Mezi webovým klientem Krameria a DL4DH Feederem lze snadno přecházet pomocí tlačítka v záhlaví obou webů.

Pro plnohodnotnou práci s Feederem je nutné, aby se uživatel přihlásil, čímž získá jednak přístup k většímu množství funkcí, ale také k evidenci historie dotazů umožňující zjistit, zda v nastavení Feederu nebo v podkladových datech nedošlo od doby, kdy s Feederem naposledy pracoval, k nějaké změně. Přihlásí-li se badatel z výzkumné instituce prostřednictvím účtu své organizace, bude mít pro potřeby vytěžování textů a dat přístup nejen k dílům volným, ale i k dílům jinak chráněným autorským zákonem (viz kapitola 2.5). Při přihlášení musí uživatel souhlasit s podmínkami použití systému, v nichž jsou vymezena práva a povinnosti uživatelů.

Úvodní obrazovka DL4DH Feederu nabízí jednoduché vyhledávání v dokumentech a datech obsažených v systému Kramerius+ a v propojené digitální knihovně Kramerius. K vyhledávání slouží vyhledávací řádek uprostřed úvodní obrazovky. Zaškrtávací tlačítko „pouze veřejné“ omezuje hledání na tzv. volná díla. Vyhledávací řádek obsahuje nápovědu v podobě našeptávače nabízejícího dostupné tituly podle názvu. Zadané výrazy se vyhledávají v názvech dostupných dokumentů, jejich plných textech i metadatech. Do Feederu je také možné pouze vstoupit

[98] Tj. zakončené na /swagger-ui/index.html?configUrl=/v3/api-docs/swagger-config.
[99] Zdrojový kód: https://github.com/ceskaexpedice/kramerius-web-client.

a jeho obsah procházet, například pomocí připravených filtrů. V takovém případě je možné nechat vyhledávací řádek prázdný a kliknout na tlačítko „Vstoupit do DL4DH Feederu“.



Obrázek č. 5 Úvodní obrazovka DL4DH Feederu

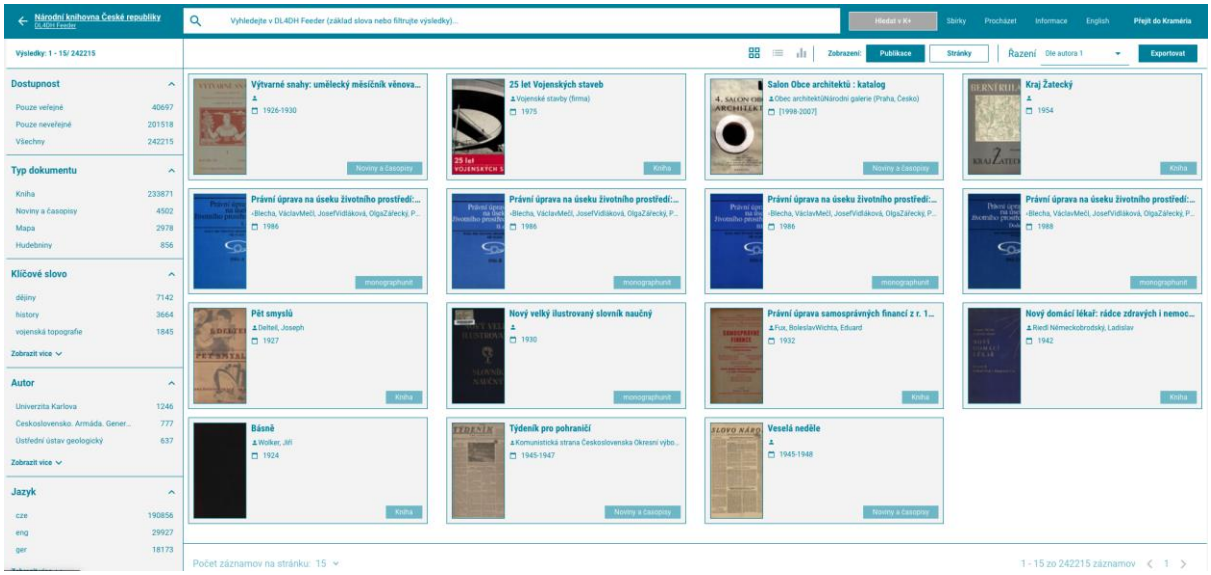
V pravém horním rohu úvodní obrazovky se nachází nabídka pro změnu uživatelského rozhraní na anglické, zobrazení informací o projektu DL4DH a pro přechod na přihlašovací stránku. Na přihlašovací stránce je vidět seznam organizací, s jejichž institucionálním účtem je možné se přihlásit.

Po vstupu do DL4DH Feederu se uživatel dostane na obrazovku s výsledky vyhledávání, případně s veškerým obsahem Feederu, pokud nebyl žádný vyhledávací výraz zadán. Vyhledávací pole v horní části obrazovky se využije při hledání dokumentů obsahujících rozpoznané entity identifikované ve fázi obohacování (např. geografické údaje, instituce či osoby); při zápisu podmínky našeptávač zobrazuje v tomto poli hodnoty, které jsou relevantní pro aktuálně vyfiltrovanou sadu dokumentů. Pokročilé vyhledávání slouží k formulování dotazu, který kombinuje více prohledávaných atributů, logické operátory a zástupné znaky.

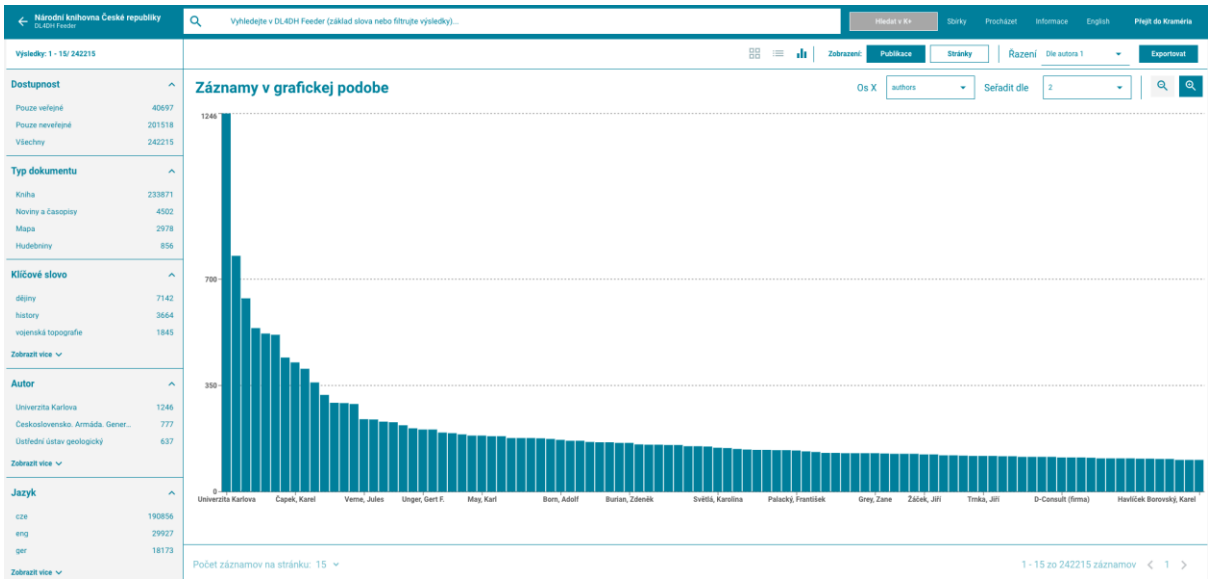
V levé části obrazovky se nacházejí fasety určené ke zpřesnění výsledku vyhledávání. Filtrovat je možné podle dostupnosti dokumentů, typu dokumentu (knihy, noviny a časopisy, hudebniny…), klíčových slov, autorů, geografických názvů, ale také podle dat získaných obohacením publikací (podle technických údajů o obrazovém dokumentu, softwaru použitém při digitalizaci aj.). Položky nabízené v rámci jednotlivých faset vycházejí z metadat o dokumentech v systémech Kramerius+ a Kramerius, a závisejí tedy na konkrétní digitální knihovně, nad níž jsou

nástroje DL4DH nasazeny. Použité filtry omezující současné vyhledávání se řadí na začátek levého sloupce s fasetami. Kliknutím na konkrétní filtr ho lze zrušit. Na kombinaci dotazu a aplikovaných filtrů reaguje seznam publikací v hlavní části obrazovky. Nalezené dokumenty je potom možné řadit podle relevance či data vydání nebo abecedně podle titulu, autora apod.

Ve výchozím nastavení jsou jednotlivé dokumenty prezentovány ve formě karet nabízejících základní bibliografické informace a náhledovou stránku dokumentu. Zobrazení je možné v záhlaví obrazovky přepnout do podoby tabulky nebo grafu, u něhož lze volit hodnoty vynesené na horizontální ose. Výsledky hledání lze zobrazit po celých publikacích nebo jednotlivých stranách s výskytem hledaného výrazu či pojmu. Kromě přepínání zobrazení je v horní části obrazovky dostupné také řazení vyhledaných položek a tlačítko pro export dat. Na jedné obrazovce se zobrazuje pouze omezené množství výsledků, k dalším je možné se dostat listováním pomocí šipek v pravém dolním rohu obrazovky.



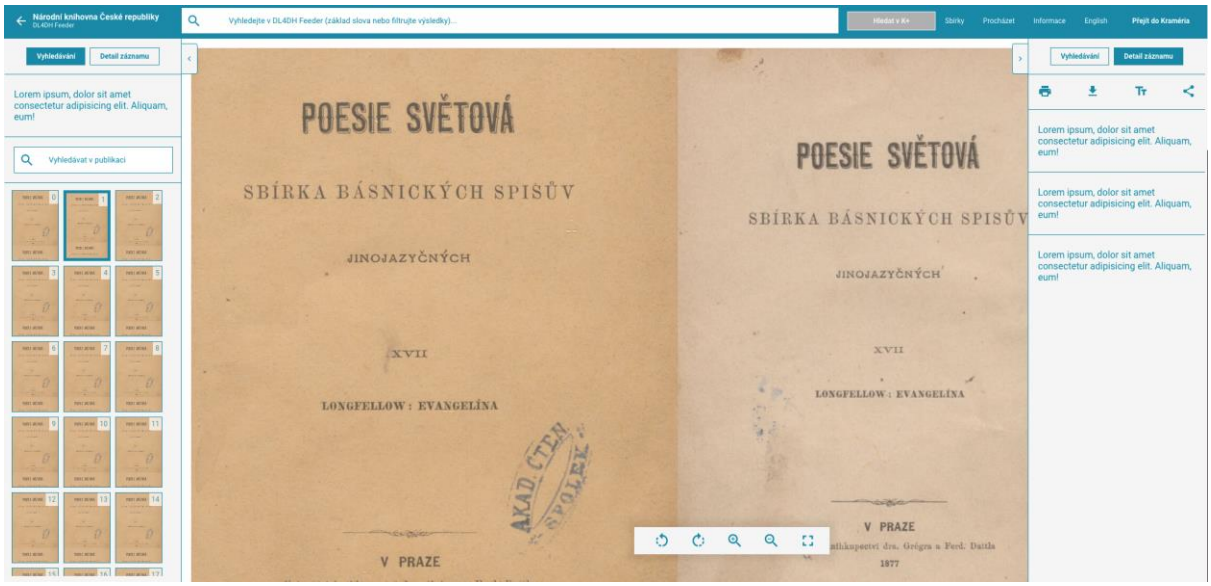
Obrázek č. 6 Obrazovka s výsledky vyhledávání



Obrázek č. 7 Výsledky vyhledávání vyjádřené grafem

Pro zobrazení podrobností o dokumentu a obrazových dat slouží kliknutí na požadovaný dokument ve výsledcích vyhledávání. Po levé straně se nachází výpis stránek dokumentu a nad ním vyhledávací řádek pro hledání v jeho plném textu. Levá část obrazovky slouží k přepínání mezi vyhledáváním v dokumentu a detaily o dokumentu. Kliknutím na vyčnívající pole s šipkou vpravo je možné levý postranní sloupec skrýt.

Dokument je možné vytisknout, sdílet nebo stáhnout včetně souvisejících metadat. Po najetí kurzorem do dolní části obrazovky se ukáže nabídka s dalšími možnosti práce s dokumentem, jako je zvětšování, zmenšování, otočení stránky nebo roztažení stránky přes celou obrazovku. Pravou část obrazovky je také možné přepnout do režimu porovnávání a prohlížet si dva dokumenty vedle sebe. Při paralelním prohlížení dvou dokumentů se na pravé straně nacházejí podrobnosti o druhé publikaci. V prezentační části webové stránky lze zvolit i jiné způsoby zobrazení obsahu: dvě (po sobě jdoucí) strany téhož dokumentu ve formě obrázku, popř. rozpoznaného textu a kombinaci obrázku strany a rozpoznaného textu na téže stránce.



Obrázek č. 8 Režim porovnávání dvou dokumentů

Nalezené publikace, popř. jejich vybrané strany lze exportovat a stáhnout na zařízení badatele kvůli dalšímu zpracování. Po kliknutí na odpovídající tlačítko si uživatel v dialogu zvolí cílový formát exportu (TEXT, CSV/TSV, JSON, TEI nebo ALTO) a nastaví požadované parametry, které se u jednotlivých formátů částečně liší. Nedojde-li k omezení exportovaných částí, obsahují exportovaná data maximum dat, metadat i paradat získaných jak z původní digitální knihovny, tak v procesu obohacování.

7.2.2 REST API

Badatel se musí nejprve autentifikovat, tj. přihlásit prostřednictvím volání `/api/login`. Pro vyhledání relevantních publikací slouží koncový bod `/api/search`.

Koncové body na adrese `/api/publications/{id}` slouží k získání přehledu o publikacích, popř. jejich součástí (čísel u periodik, konkrétních stran) dostupných v systému Kramerius+.

Koncové body na adrese `/api/exports/` slouží pro získání přehledu o exportovaných publikacích (`/api/exports/list`), pro spuštění procesu exportování s požadovanými parametry (`/api/exports/{id}/csv` apod.) nebo ke stažení souboru, který na základě exportu publikace vznikl (`/api/exports/download/{id}`).

Podrobnější a aktuální informace o programovém rozhraní REST API jsou dostupné v úložišti zdrojového kódu, popř. z URL adresy pro dokumentaci pomocí OpenAPI[100].

[100] Tj. zakončené na /swagger-ui/index.html?configUrl=/v3/api-docs/swagger-config.

7.3 Známá omezení a problémy

Obohacování dat naráží na některá úskalí, která projekt DL4DH v současné době neřeší. V prvé řadě je to nízká kvalita vstupních dat, tj. digitálních obrazů publikací, případně různá kvalita automatického rozpoznání textu[101]. To může být příčinou chyb při následných analýzách lingvistickými nástroji.

Na práci lingvistických nástrojů má dále vliv kombinace několika jazyků v textu. I když používané nástroje pro morfologickou analýzu (UDPipe 2), resp. pro rozpoznání entit (NameTag 2), umějí analyzovat několik jazyků, správná aplikace vyžaduje nejprve identifikaci pasáží s odlišnými jazyky a jejich samostatné zpracování s odpovídajícím nastavením parametrů jednotlivých programů.

Uložené dotazy, popř. data, jež těmto dotazům odpovídají, nelze považovat za stoprocentně replikovatelná, neboť může dojít k aktualizaci bibliografických metadat, k novému rozpoznání pomocí OCR, případně k analýze textu novějšími verzemi lingvistických nástrojů, přičemž ani v jednom z uvedených případů se předchozí verze dat neuchovávají.

[101] Na kvalitu OCR může mít vliv např. software, který nedokáže zpracovat text obrácený o 90 stupňů, nebo volba nesprávného modelu (typu písma) pro rozpoznávání.

8 Využití nástrojů DL4DH ve výzkumné praxi

Materiál dostupný v digitálních knihovnách je velmi bohatý a různorodý. Stejně jako existuje velké množství badatelských otázek, které čekají na odpověď, existuje i mnoho způsobů, jak data z digitálních knihoven uplatnit, přičemž o některých v této chvíli možná ani nevíme. Následující kapitola přináší soubor příkladů ilustrujících celý proces vědecké práce, položenými otázkami počínaje a prezentací výsledků bádání konče. Výběrově jsou zde uvedeny scénáře z několika humanitních oborů, které naznačují, k čemu je možné v praxi využít data a nástroje, které vznikly v rámci projektu DL4DH.

8.1 Biblické citáty v periodickém tisku

Jedním z možných přístupů k výzkumu společenské relevance biblických textů je sledovat užití biblických citátů v rámci periodického tisku. S tiskem přichází denně do styku značné množství lidí a biblické citáty zde mohou často fungovat jako významný faktor symbolické komunikace. Jako základ studie, která zkoumá jednotlivé významové roviny biblických citací, je potřeba nejprve biblické citáty v tisku identifikovat. Vzhledem k rozsahu materiálu (jak bible samotné, tak zkoumaných periodik) je tento úkol běžnými metodami velmi zdlouhavý. Automatické zpracování dat tak poskytuje badatelům výraznou pomoc.

8.1.1 Výzkumná otázka

Základní výzkumná otázka zní: Jaké jsou vzorce citování biblického textu v periodickém českojazyčném tisku? Cíle této studie jsou zejména statistické a připravují data pro další analýzy.

8.1.2 Materiál

Studie předpokládá práci s plnými texty vybraných periodik, z nichž je možné vytvořit uživatelskou datovou sadu a exportovat potřebná data pomocí DL4DH Feederu. Biblický materiál sestává z několika českých překladů, u kterých očekáváme užívání ve vytyčeném období.[102]

[102] V rámci projektu byly využity překlady Bible Kralická, Nový zákon (Dr. František Žilka), Starý zákon a deuterokanonické knihy (Dr. Jan Hejčl), Nový zákon (Dr. Jan Ladislav Sýkora), které v přepisu poskytl Libor Diviš (Studijní on-line bible). Rozsáhlé části bible v překladu Bible Svatováclavská poskytl prof. Pavel Kosek z Ústavu českého jazyka Masarykovy univerzity. V budoucnu není vyloučené rozšíření o další překlady.

8.1.3 Metoda

V DL4DH Feederu vybereme periodika, ve kterých chceme biblické citáty hledat, a exportujeme je jako prostý text spolu s metadaty (potřebné údaje jsou: název periodika, datum vydání a UUID jakožto identifikátor vydání).[103]

Celý proces hledání biblických citací je rozdělen do několika fází, které musí reflektovat zejména obecně problematický stav prohledávaného korpusu (stav OCR a používání různých překladů biblického textu, odlišných od srovnávacího vzorku). Toto prohledávání se již odehrává v rámci specifického postupu u badatele, tedy mimo DL4DH Feeder.

1) Předzpracování biblického korpusu:
    a) Rozdělení jednotlivých veršů na jejich relevantní menší části (tzv. podverše).
    b) Vektorizace podveršů na základě n-gramů o velikosti nejvýše 4 znaky a příprava slovníku n-gramů pro následnou vektorizaci prohledávaného souboru dat.
2) Vyhledávání ve zvoleném korpusu:
    a) První fáze vyhledávání je založena na připravených vektorech. V menších částech prohledávaného korpusu (o velikosti cca 6 vět) hledáme alespoň 85% shodu (z počtu n-gramů podverše) s jednotlivými vektory podveršů.
    b) Pokud je nalezena vektorová shoda, pasáž je ověřena pomocí tzv. Levenštejnovy (editační) vzdálenosti (dále LD).[104] Potřebná shoda je v tomto případě nastavena na 85 % (z počtu znaků podverše).
    c) Takto nalezené výsledky jsou uloženy do samostatného souboru ve formátu CSV pro další kontrolu.
3) Automatizovaná kontrola detekovaných biblických citací, která má za cíl vypořádat se s následujícími jevy:
    a) Vyřešení duplicitních výsledků (např. verš byl detekován ve více překladech nebo podverších).

[103] Detailní popis celého postupu vyhledávání biblických citací je popsán na adrese https://dl4dh.nkp.cz. Jádro projektu je vystavěno v jazyku Python (verze 3.9).
[104] Viz https://en.wikipedia.org/wiki/Levenshtein_distance [cit. 3. 5. 2022]. Ve stručnosti „kolik znaků se musí v jednom textu změnit, aby byl stejný jako jiný text“.

b) Vyřešení vícečetné atribuce v rámci jedné pasáže (např. pasáž skutečně obsahuje více citací vs. verše/podverše jsou si velice podobné).

c) Predikci pravděpodobnosti, že detekovaná pasáž skutečně obsahuje biblický verš a ne jen něco, co je mu podobné (zjišťováno např. na základě LD či celkového počtu detekovaných podveršů).

d) Zkontrolované výsledky jsou uloženy do samostatného souboru ve formátu CSV pro závěrečnou ruční kontrolu.

4) Ruční kontrola výsledků, která je nutná kvůli mnohým problémům, které plynou z povahy prohledávaného i biblického korpusu.

8.1.4 Předpokládaný výsledek

Zvolené řešení by mělo vést k detekci biblických citací ve zvoleném korpusu. Statistické výstupy studie je možné následně vizualizovat v grafech, které znázorňují vzorce citování biblického textu v závislosti na jednotlivých periodikách i na vývoji v letech. Tyto statistiky mohou také být výchozím bodem pro sledování vztahů mezi jednotlivými periodiky založené na základě citovaných biblických veršů. Stejně tak mohou být sledovány vztahy mezi verši na základě toho, v jakých periodikách se objevují. Tyto vztahy mohou být vyjádřeny například pomocí síťových grafů. V neposlední řadě je možné dále pracovat s jednotlivými citacemi a zkoumat je v kontextu periodik.

Tato studie byla v praxi realizována jako pilotní test v rámci projektu DL4DH na vzorku periodik z let 1925 až 1939. Kompletní výstupy a popis projektu je dostupný na adrese https://dl4dh.nkp.cz, finální výstupy jsou k dispozici od září 2022. Statistické výstupy i grafy jsou koncipovány jako otevřené zdroje a badatelé je mohou dále analyzovat. Volně dostupným výstupem projektu je též samotný algoritmus[105]. Celý postup je tak možno reprodukovat nad libovolnou kolekcí nebo uživatelskou sadou dat.

8.2 Archeologické lokality v historickém místopisu

Primárními prameny pro archeologický výzkum jsou archeologické nálezy identifikované při terénních výzkumech. Dochování nálezů v krajině je však závislé na řadě okolností, včetně ča-

[105] Viz https://github.com/DigilabNLCR/BibleCitations.

sového odstupu od jejich uložení do současnosti. Významný vliv hraje způsob využití dotčeného území, který ovlivňuje možnost dochování archeologických nálezů – nálezy mohou být zasaženy zemědělskou činností, výstavbou, přirozenou erozí apod. Zintenzivnění dopadů lidských aktivit na krajinu v průběhu 20. století vedlo ke zničení velkého množství dříve dobře dochovaných archeologických památek, které dnes již nedokážeme v terénu identifikovat. Tyto památky však mohou být zachyceny a popsány ve starších publikovaných zdrojích, zejména v historických místopisech, které postihují krajinu v době před razantní urbanizací a změnou struktury zemědělské krajiny ve druhé polovině 20. století. Údaje o archeologických lokalitách bývají z místopisných prací vytěžovány dosud pouze namátkově, neboť jde o časově náročnou práci. Pomocí nástrojů DL4DH lze k tématu přistoupit systematičtěji a efektivněji, s cílem evidovat dnes již zapomenuté archeologické památky.

8.2.1 Výzkumná otázka

Jak často starší historicko-místopisné práce obsahují informace o existenci archeologických památek? Jak často jde o lokality, které dnes nejsou součástí evidenčních seznamů? Dokážeme identifikovat archeologické lokality, kde došlo k významné změně stavu dochování? O jakých typech nálezů se místopisy obvykle zmiňují a které zachyceny zpravidla nejsou?

8.2.2 Materiál

Jako součást digitální knihovny NK ČR byla sestavena speciální kurátorská kolekce místopisných prací, postihující několik stovek základních děl z oblasti historického místopisu. Do kolekce jsou zařazeny především nadregionální práce, systematicky postihující větší území. Byť celkový počet dosud vydaných místopisných prací přesahuje 7 tisíc titulů (údaj vychází z rešerše provedené na portálu Bibliografie dějin Českých zemí[106], skutečné číslo může být ještě vyšší), jejich užití je omezeno stavem digitálního zpracování a zpřístupnění v jednotlivých knihovnách. V případě komplexního zpracování by bylo žádoucí využít všechna místopisná díla, pro základní přehled však dobře poslouží alespoň sbírka zahrnující základní tituly, jako jsou Soupis památek uměleckých a historických, jednotlivé práce A. Sedláčka, místopisy G. Sommera, J. Schallera a dalších autorů.

[106] https://biblio.hiu.cas.cz

8.2.3 Metoda

1) V kolekci digitálních dokumentů pomocí DL4DH Feederu vyhledáme výskyt pojmů naznačujících přítomnost archeologických památek v terénu (např. hradiště, mohylník, hrad, tvrz, zaniklá vesnice apod.). Pro výběr vhodných pojmů lze využít Tezaurus archeologické terminologie[107].

2) Relevantní strany publikací exportujeme ve formátu CSV. Z exportu lze vyloučit morfosyntaktické údaje jednotlivých tokenů, které pro další kroky zpracování nejsou relevantní. V mnoha případech může být vhodné exportovat celé publikace a výslednou pracovní datovou sadu vybírat až na základě exportu, neboť lokační údaje a věcný popis se nemusí vždy nacházet na téže tiskové straně.

3) V exportovaných datech odstraníme nadbytečné řádky, tj. takové části textů, které jsou dostatečně vzdálené od hledaných pojmů. Cílem těchto úprav je zachovat nalezené lokační údaje (místní a pomístní jména) s relevantním kontextem v textu.

4) Výsledky seskupíme tak, aby každá nalezená zmínka měla přiřazenou geografickou lokalizaci. Na tomto základě sestavíme geoprostorovu databázi.

5) Provedeme srovnání databáze s existující archeologickou evidencí, získanou např. prostřednictvím Digitálního archivu Archeologické mapy České republiky[108].

8.2.4 Předpokládaný výsledek

Uvedeným postupem bude vytvořena srovnávací databáze prostorově lokalizovaných zmínek o vybraných typech archeologický lokalit s vazbou na původní zdroj (místopisnou publikaci) a na současnou evidenci archeologických lokalit. Databáze umožní vyhodnotit četnost, typové zastoupení a informační hodnotu zmínek o archeologických památkách v jednotlivých historických místopisných zdrojích.

8.3 Identifikace veršů v digitalizovaných publikacích

Básnický jazyk se od běžné mluvy odlišuje používáním uměleckých prostředků (obraznost nebo symbolika). Poezie se od prózy odlišuje používáním rýmu, rytmu nebo metra. Básnictví jakožto jedna z forem literatury zaujímá v jazyce a jeho výzkumu speciální postavení, viz např.

[107] https://teater.aiscr.cz
[108] https://digiarchiv.aiscr.cz

versologický tým Ústavu pro českou literaturu[109] a Korpus českého verše[110]. Specifická podoba grafického záznamu veršovaného textu (rozdělení na verše a strofy) se odráží také v prostředcích, které se pro jeho zachycení používají, viz např. samostatná kapitola ve standardu TEI[111]. Výstupy z programů pro rozpoznávaní znaků (OCR) rozdíly mezi prozaickým a veršovaným textem nesignalizují. Identifikace textového řádku jakožto verše je ponechána na dalším zpracování výstupu z těchto aplikací.

8.3.1 Výzkumná otázka

Cílem studie je pomocí analýzy výstupů OCR ve formátu ALTO identifikovat formální prvky, které umožní označit řádek textu jako (básnický) verš. K těmto prvkům může patřit zarovnání, popř. odsazení bloku textu, šířka a pozice řádku vůči předchozímu a/nebo následujícímu.

8.3.2 Materiál

Vytvoříme dva vzorky publikací (uživatelské datové sady): veršované a prozaické texty. V kolekci pomocí DL4DH Feederu vyhledáme básnické publikace pomocí pokročilého dotazu (titul obsahuje text „Básně“). Následně pomocí faset vybereme pouze knihy v češtině. Pro nalezené publikace uložíme výstupy ve formátu ALTO. Získaný materiál rozdělíme na dvě části: analytickou (5 %) a testovací (95 %). Pro vzorek prozaických titulů vybereme publikace, které s velkou pravděpodobností veršované texty neobsahují, a to pomocí klíčových slov jako např. „Dějiny“, „Mezinárodní vztahy“, „Politika a vláda“ apod. V rámci příprav dat odstraníme u každé publikace několik počátečních a koncových stran, které s velkou pravděpodobností budou obsahovat neveršovaný text, a to i u básnických titulů. Následně vytvoříme vzorky o přibližně stejném počtu stran (tj. dokumentů ve formátu ALTO).

8.3.3 Metoda

Texty ve formátu ALTO uložíme kvůli rychlejšímu zpracování ve specializované databázi, např. eXist-db[112] nebo BaseX[113]. Formální a automatizovanou analýzou analytické části vzorku určíme elementy, jejich atributy a hodnoty, které umožní identifikaci řádku s veršem. Těmito parametry budou např. rozměry stránky (atributy @HEIGHT a @WIDTH elementu <Page>),

[109] https://versologie.cz
[110] https://versologie.cz/v2/web_content/corpus.php?lang=cz
[111] https://tei-c.org/release/doc/tei-p5-doc/en/html/VE.html
[112] https://exist-db.org
[113] https://basex.org

zarovnání bloku textu, odsazení textového řádku (atribut @VPOS elementu <TextLine>). Vodítkem může být také počet slov na řádku (resp. elementů <String>) a zakončení řádku identickou skupinou písmen (rýmy).

V programovacím jazyce XQuery[114] vznikne program, jenž pomocí vlastního atributu v rámci jedné stránky ve formátu ALTO označí elementy <TextLine>, které s velkou mírou pravděpodobnosti tvoří verš. Pomocí propojení formátu ALTO se serverem IIIF dojde ke zvýraznění identifikovaných veršů v digitalizované kopii, které poslouží k vizuální kontrole správného označení verše.

Program se aplikuje na testovací sadu dat, a to na texty básnické i prozaické. V druhém vzorku by se verše neměly vyskytovat. Případný výskyt verše je potřeba ověřit a v případě falešných výsledků upravit algoritmus programu.

8.3.4 Předpokládaný výsledek

Výsledkem studie bude algoritmus, který na základě nastavitelných parametrů dokáže identifikovat a ve formátu ALTO označit verše v básnické skladbě. Tato informace bude využitelná při transformaci textu do jiných strukturovaných formátů, např. TEI.

8.4 Hodnocení dějin českého knihtisku

Pohled na národní knižní kulturu byl v minulosti – podobně jako historiografie – poznamenám ideologickou zátěží. V 19. století byl obraz národní knižní kultury konstruován pod vlivem obrozeneckého národního étosu, v období první republiky pod vlivem protestantských historiků. Konečně v době po druhé světové válce byl formován v souladu s marxistickým společensko-politickým diskurzem. V rámci hodnocení českých dějin knihtisku se proto největší pozornosti těšil bratrský knihtisk, přestože objemem představoval zanedbatelnou část tiskařské produkce. Tradičně se soudí, že ke glorifikaci tištěné produkce pronásledované minoritní náboženské komunity došlo zejména v období první republiky pod vlivem protestantských historiků Františka Palackého a T. G. Masaryka. Avšak i poválečné období, podobně jako pozdější normalizační ideologie adorující ideje husitství, mohlo výrazně ovlivnit náš pohled na starší dějiny knižní kultury, v nichž přední místo zaujala tištěná produkce tiskařské dílny Jednoty bratrské.

[114] https://www.w3.org/TR/xquery-31/

8.4.1 Výzkumné otázky

V jakém typu periodického tisku, popř. v jakém kontextu byly publikovány zprávy týkající se starších dějin knižní kultury (např. vydavatelských domů, dějin knihtisku, nejstarší domácí tištěné publikace apod.) ve 20. století? Jaké knihovědné práce (autoři/témata) byly nejvíce citované v období první republiky a po roce 1948?

8.4.2 Materiál

Vytvoříme na základě filtrace dat, formátů a názvů čtyři sady vzorků (uživatelských datových sad):

- denní tisk z období první republiky (1918–1938)
- odborné texty z období první republiky (1918–1938)
- denní tisk z období po druhé světové válce (např. 1950–1970)
- odborné texty z období po druhé světové válce (např. 1950–1970)

8.4.3 Metoda

V plných textech periodik ze získaných vzorků vyhledáváme klíčová slova týkající se knižní kultury a dějin knihtisku (např. knihtisk, knihtiskař, knihověda, knižní kultura, tiskařský lis, Jednota bratrská, Melantrich, Bible apod.). Exportujeme text stran s nalezenými výrazy, a to ve formátu CSV včetně lemmat. Vytvoříme uživatelské sbírky s výskyty klíčového slova jako výchozí srovnávací materiál pro další komparativní průzkum. Ke každému výskytu bude k dispozici také seznam lemmat v bezprostředním okolí výrazu. Sledujeme kontext, v jakém jsou užita. Všímáme si, v jaké rubrice denního periodika se klíčová slova vyskytují, sledujeme též slova a slovní výrazy užité ve spojitosti s klíčovými slovy. Vznikne databáze s nalezenými prvky (tj. s klíčovými slovy, resp. odkazy na odbornou literaturu), která bude obsahovat další údaje pro ověření hypotézy (identifikace periodika, datace, rubrika, nejfrekventovanější kolokace). Pro identifikaci odkazů na odbornou literaturu využijeme nástroj Grobid.

8.4.4 Předpokládaný výsledek

Navržené řešení by mělo vést k identifikaci tiskovin, které obsahovaly zprávy o dějinách knihtisku či knižní kultuře obecně, a to včetně přesnější lokalizace v rámci periodického tisku či odborných textů. Zvolený postup umožní vytvoření uživatelské sbírky se soupisem všech citací, v nichž se hledaná klíčová slova včetně kolokací vyskytují. Soupis citací bude výchozím materiálem při sledování společenského a ideologického kontextu, v jehož rámci byly dějiny starší

knižní kultury tematizovány v době první republiky a v poválečném období. S využitím aplikace Grobid bude možné vytvořit soupis citované odborné literatury ve sledovaných obdobích 20. století

8.5 Proměny vědecké komunikace z perspektivy sociologie vědy

V klasické práci sociologie vědy popsal Price (1963) vědu jako systém, který se rozvíjí po trajektorii exponenciálního růstu, a předpovídal jeho saturaci. Expanze vědy v éře tzv. velké vědy (big science) vede ke změnám v podobě vědecké práce. Spoluautorství publikací se stává normou, snižuje se průměrná kvalita vědeckého personálu, narůstají náklady, zdroje se koncentrují v unikátních infrastrukturách a dochází k těsnějšímu sepětí vědy a politické sféry. Proměny vědecké praxe vyvolávají otázku dopadu těchto změn na vědecké poznání. Stává se fragmentovanějším a specializovaným? Vede zvýšená konkurence k nedostatečné replikovatelnosti vědeckého výzkumu? Jsou publikované výsledky spíše inkrementálními příspěvky k dosaženému poznání?

Povahu vědeckého poznání je obtížné měřit. Můžeme však vyjít z premisy, že různé druhy poznání je nutné komunikovat pomocí odpovídajících jazykových prostředků. Jako materiál pro výzkum tak může sloužit podoba vědecké komunikace zachycená v odborných textech publikovaných především v akademických časopisech. Digitální knihovny, které obsahují a zpřístupňují odborná periodika, poskytují vhodné prostředí pro zkoumání vědecké komunikace z perspektivy sociologie vědy.

8.5.1 Výzkumná otázka

- Jak se proměnila vědecká komunikace v čase?
- Vyvíjí se vědecká komunikace rychleji než běžný jazyk?
- Mění se vědecké poznání v souvislosti se změnou vědecké práce?

8.5.2 Materiál

Koncept výzkumu vyžaduje vytvoření čtyř vzorků účelovým výběrem:

- odborné texty ze staršího období (např. 1870–1880)
- odborné texty z mladšího období (např. 1970–1980)
- denní tisk ze staršího období (např. 1870–1880)
- denní tisk z mladšího období (např. 1970–1980)

Takto sestavený vzorek umožňuje realizovat výzkum, v němž existují dva páry dat pro experimentální a kontrolní skupinu. Odborná periodika představují dominantní platformu pro publikaci vědeckých výsledků, lze na nich tedy sledovat vývoj vědecké komunikace. U denního tisku lze předpokládat, že kopíruje obecný vývoj jazyka. Funkce denního tisku lze považovat za dějinně relativně konstantní, může tedy sloužit jako srovnávací kritérium. Avšak na místě je opatrnost: například česká jazyková data z 2. poloviny 20. století mohou být natolik ovlivněna propagandistickými funkcemi denního tisku, že se od obecného jazyka odklánějí.

8.5.3 Metoda

Podkladová data získáme z DL4DH Feederu pomocí metadat, resp. faset, kdy „rok vydání“ vymezuje časové období pro každý vzorek, „typ dokumentu“ omezí vyhledávání na periodika, „jazyk“ vyloučí dokumenty v jiném než českém jazyce a „klíčová slova“ zajistí oborovou srovnatelnost v rámci vzorku odborných textů. Porovnat je tak např. možné pod klíčovým slovem „Matematika“ vybrané ročníky *Časopisu pro pěstování mathematiky a fysiky* vycházející od konce 19. století do poloviny 20. století a na něj navazující *Časopis pro pěstování matematiky* vycházející ve 2. polovině 20. století. Podobným způsobem sestavíme kontrolní vzorek z denního tisku, který by měl svým rozsahem a časovým obdobím odpovídat experimentálnímu vzorku. Finální vyhledávací výraz je žádoucí uložit pro případné opakované použití.

Vyhledaná data stáhneme ve formátu CSV v podobě obohacené o morfologickou analýzu a detekci vlastních jmen. Příslušné sloupce umožní vyfiltrovat pouze významová a funkční slova a vyloučit interpunkci i vlastní jména. Pro analýzu pak vybereme sloupec dat obsahující lemmata.

Kvalitu textu v datech je vhodné manuálně kontrolovat na náhodných vzorcích. Informace z paradat o použitém softwaru pro optické rozpoznání znaků (OCR), která DL4DH Feeder poskytuje, mohou poskytnout zpětnou vazbu v případě, že mezi různými publikacemi existují systematické rozdíly.

Ze získaných dat můžeme vypočítat statistiky, u nichž předpokládáme, že indikují námi sledované jevy. Statistiky můžeme normalizovat např. tak, že každý vzorek rozdělíme na úseky o stejném počtu slov, která vybereme náhodným výběrem bez opakování. Tímto způsobem pak lze v rámci pozorování specializace vědeckých textů spočítat třeba průměrný počet unikátních slov na každých 1000 náhodně vybraných slov. Konstrukce relevantních indikátorů představuje

nejnáročnější úkol, při kterém musíme uplatnit znalost předchozí odborné literatury o vědecké komunikaci.

Posledním krokem analýzy je vyhodnocení statistik. V našem návrhu výzkumu není primárně důležitá velikost rozdílu mezi indikátory odborného a běžného textu v rámci jednoho časového období, ale sledujeme především to, zda se velikost rozdílů zvýšila v čase. Běžný text tedy slouží jako srovnávací měřítko, které mj. kontroluje efekt jazykového vývoje. Pokud však zjistíme výrazné změny i v běžném textu, může to být signálem, že tento ukazatel neplní svou funkci.

8.5.4 Předpokládaný výsledek

Analýza ukáže, zda se historické proměny vědy jako sociální instituce odráží i ve způsobu vědecké komunikace. Zatímco existuje dostatek důkazů o změnách v organizaci vědecké práce, není zřejmé, do jaké míry mají tyto změny dopad na povahu vědeckého poznání. V případě, že detekujeme významné posuny ve většině sledovaných indikátorů, a naše hypotézy tak budou podpořeny, získáme empirickou evidenci pro tvrzení, že výzkumná praxe a její výsledky spolu souvisí a že obojí se v průběhu času proměnilo. V opačném případě dosáhneme tzv. nulového výsledku. Pokud jsme postupovali správně a systematicky a zvolili vhodné indikátory, je možné i takový výsledek publikovat, zejména pokud je v protikladu s empiricky nepodloženými tvrzeními v literatuře. Při interpretaci výsledků musíme dbát na omezení našeho výzkumu, který se zabýval jen některými obory (různé vědecké disciplíny se vyvíjejí odlišným tempem) a je národně a jazykově specifický.

8.6 Makroanalýza jazyka a témat české beletrie

Literatura a jazyk se dají studovat pomoci rozličných přístupů. Zatímco zejména v literární vědě dosud převládají kvalitativní přístupy ve smyslu „blízkého čtení“ (close reading), nelze nepřihlédnout i ke kvantitativním metodám, které se zabývají se možnostmi „vzdáleného čtení“ (distant reading). Tyto postupy mohou exaktně zpracovávat velké množství textu a pomocí makronalýzy přispět k poznání nebo potvrzení jevů, které čtením zblízka nelze jednoznačně uchopit. Obdobný postup nastiňuje Jockers (2013). Výstavba rozsáhlých repozitářů národní li-

terární tvorby spolu s automatizovaným exportem dat výzkum metodami vzdáleného čtení podporuje. Tento postup dává důraz na integraci rozsáhlého korpusu, společně s maximální automatizaci jeho správy a návazných analytických postupů nad rozsáhlou sadou dat.

Základním předpokladem tohoto výzkumu je tvrzení, že literaturu lze hodnotit v kontextu své agregované minulosti a zároveň budoucnosti, která po ní následovala. Je-li digitalizováno výrazné množství publikací, je možné, s vědomím chybějících částí, predikovat širší závěry. Na literaturu je však možné nahlédnout prizmatem kulturních a sociálních vztahů, které jsou v našem případě dokumentované ve stejném repozitáři, a to prostřednictvím tiskových a vědeckých médií, které pro 19. a většinu 20. století pokrývají velkou část mediálního prostoru psaného jazyka.[115]

8.6.1 Výzkumná otázka

Lze přiřadit jednotlivým (níže vybraným) synchronním a diachronním segmentům beletrie specifické užití jazyka (lexikální, morfologické, syntaktické), tematických kombinací a sémantických vzorců? Lze tato užití a jejich statistickou významnost originálních a statisticky významných kombinací vysledovat v rámci proudu diseminace v beletrii? Lze tyto změny korelovat s dalšími druhy a kanály diseminace textů?

8.6.2 Materiál

Pro správu rozsáhlého korpusového souboru nelze využít ukládání uživatelské datové sady jako primárního nástroje DL4DH Feederu, ale korpusový manažer a vlastní databázi (místní nebo SaaS, tj. software jako služba). Rozsah jednotlivých vzorků je možné vymezit jak v synchronní perspektivě (na základě kategorizace rozsahu, předmětových hesel), tak diachronně na základě kulturně společensky významných období, nebo vývoje a diseminace jednotlivých literárních žánrů a hnutí.

8.6.3 Metoda

Tento postup využívá vlastní instanci korpusového manažeru Manatee[116] a uživatelského klientu, např. opensourcového KonTextu[117], nebo lokální instanci Sketch Engine[118]. Pomocí we-

[115] Pro novější období bude nepochybně nutné zohlednit také nová média a rozličné internetové zdroje.
[116] https://nlp.fi.muni.cz/trac/noske
[117] https://github.com/czcorpus/kontext
[118] https://www.sketchengine.eu/documentation/local-installations/

bového, případně programového rozhraní DL4DH Feederu si badatel postupně připraví (vyfiltruje) dokumenty, které si v rámci předzpracování zaznamená do své databáze. Dokumentaci k REST API, která je zpracována pomocí knihovny Swagger UI[119], lze využít k výběru vhodných koncových bodů, jejichž volání bude součástí dlouhotrvajícího integračního procesu. Velké množství cílových dat z plánované uživatelské datové sady si rozdělí na menší části, které se budou pomocí REST API stahovat průběžně, obzvlášť u rozsáhlých periodických titulů. V rámci přípravy výzkumu je vhodné kontaktovat správce dané instance DL4DH Feederu s popisem rozsahu celé akce.

Badatel se domluví s administrátorem Krameria+ na intenzivním využívání nástrojů DL4DH. Následně si vytvoří integrační úlohu, která voláním rozhraní REST API stáhne data ve zvoleném formátu, např. JSON nebo TEI. Ideálně však konzumuje prostý text a paradata, která mohou indikovat systematické rozdíly v kvalitě OCR jednotlivých publikací.[120]

Pro jazykovou analýzu lze využít korpusový klient, případně zakomponovat volání API klientu KonText, popř. Sketch Engine[121] do vlastního zpracování včetně sumarizace výsledků. K měření tematických koncentrací je potřeba si připravit sadu samostatných analytických procesů, které budou procházet definované úseky textů a zaznamenávat koncentrace u významových slov (metodiku popisuje Čech, 2016), resp. použít dohledávání sémanticky významných sousloví a jejich srovnání pomocí vnoření slov (word embeddings)[122], např. prostřednictvím výkonné knihovny Gensim.[123] V prvním případě je nutné využít vážené průměry a další statistické prostředky pro normalizaci a možnost srovnání mezi jednotlivými texty, částmi datové sady a jejími kategoriemi. V druhém případě je nutné natrénovat model a následně ho implementovat.

Tato metoda předpokládá iterativní programový přístup, který umožní ověřit metodu a předpoklady na menších datových vzorcích a následně aplikaci celého procesu na kompletní materiál. Takto se krok za krokem (nebo pomocí optimalizačních pipelines) dosáhne zařoveň identifikace vhodných algoritmů normalizace. Při vyhodnocení jevů na základě statistické četnosti, např. pomocí procedur jazyka R, se musí počítat jak s šumem v datech, tak s nelineárností procesů

[119] https://swagger.io/tools/swagger-ui/
[120] U takto rozsáhlého případu může být vhodnější, když si obohacení dat zajistí badatel sám pomocí vlastních nebo jinak dostupných lingvistických nástrojů.
[121] Viz https://www.sketchengine.eu/documentation/api-documentation/#toggle-id-4
[122] Viz např. https://cs.wikipedia.org/wiki/Vnoření_slov
[123] https://radimrehurek.com/gensim

diseminace koncentrací, přičemž právě u této nelineárnosti mohou být identifikovány zajímavé vzory.

8.6.4 Předpokládaný výsledek

Makroanalýza může za podmínky dostatečné saturace reprezentativními daty, resp. kvalitně sestavené datové sady identifikovat zásadní témata a tematické koncentrace české beletrie v rámci synchronních i diachronních průřezů. Zároveň může poukázat na vzory a postupy diseminace témat v rámci beletrie, které jsou nosné pro jednotlivé žánry, tradice a periody. Další analýza užitého jazyka může poukázat na postupné proměny psaného projevu a korelovat ho s aktuálními poznatky. Dalším přínosem pro literární vědu bude změření tematických koncentrací v dalších typech tiskových médií a míra jejich transformativního vlivu v delším časovém horizontu na českou beletrii.

9 Závěr

Tato metodika se zakládá se na výsledcích projektu DL4DH řešeného v letech 2020-2022. Projekt zásadně posouvá výzkumné možnosti při práci s velkými daty pocházejícími z českých digitálních knihoven. Metodika je prvním uceleným metodickým návodem pro vědce a badatele v oblasti humanitních věd v Česku, který systematicky v teoretické i praktické rovině popisuje, jak lze nové možnosti digitálních knihoven využít v knihovnické a výzkumné praxi. Všechny popsané technologie a softwarové nástroje se v čase vyvíjejí, proto je třeba počítat s tím, že metodika v některých detailech postupně zastará a bude se muset po několika letech aktualizovat.

Díky spolupráci knihovníků, odborníků z oblasti digital humanities a programátorů vzniklo v rámci projektu DL4DH řešení, které nabízí badatelům z mnoha vědních oborů zcela nové možnosti práce s velkými objemy textových a obrazových dat, ale i metadat. Při jeho návrhu bylo nutné brát v potaz nejen současné technické a technologické možnosti, ale také aktuální, respektive plánovanou podobu autorského zákona, který umožní využití autorských děl při vytěžování textu a dat pro výzkumné účely.

Popisované a následně implementované řešení DL4DH je napojeno na digitální knihovnu využívající systém Kramerius, používaný všemi knihovnami zapojenými do projektu. Dostupná data a metadata se transformují a obohacují mj. pomocí nástrojů zpracování přirozeného jazyka, které vyvíjí a provozuje velká výzkumná infrastruktura LINDAT/CLARIAH-CZ. Tyto procesy mohou ovládat kurátoři digitálních sbírek prostřednictvím grafického rozhraní webové aplikace nebo voláním programového rozhraní REST API. Stejným způsobem mohou badatelé získat data pro svůj výzkum. Údaje jsou k dispozici ve formátech prostý text, ALTO, CSV/TSV, JSON a TEI, které patří k zavedeným standardům, pro něž existuje mnoho analytických nástrojů a programových knihoven. Byť jde na úrovni metodiky pouze o modelový příklad řešení, jeho implementace do praxe největších tuzemských knihoven dává metodice relevanci a ukazuje na obrovský posun, kterým výzkum v oblasti humanitních věd prochází.

Ukázky případových studií v závěru metodiky představují nepatrný zlomek toho, k čemu lze vyvinuté nástroje DL4DH (Kramerius plus, TEI Converter a DL4DH Feeder) využít. Jedná se bezpochyby o cenný příspěvek k podpoře výzkumu v digitálních humanitních vědách, který bude mít přesah i do dalších vědních oborů.

10 Seznam zkratek a vybraných pojmů

ALTO	Analyzed Layout and Text Object; formát pro popis rozvržení textu na stránce
API	Application Programming Interface; aplikační programové rozhraní
born-digital	označení dokumentů, které vznikly pouze v elektronické podobě
CC	Creative Commons; soubor veřejných licencí pro autorská díla
CSV	Comma-Separated Values; hodnoty oddělené čárkou
CSV2	Comma-Separated Values; přípona používajících pro oddělení hodnot středník
CzADH	Czech Association for Digital Humanities; Česká asociace pro digitální humanitní vědy
ČNK	Český národní korpus; velká výzkumná infrastruktura
ČR	Česká republika
DH	digital humanities, digitální humanitní vědy
DILIA	Divadelní, literární, audiovizuální agentura
DL4DH	Digital Libraries for Digital Humanities; digitální knihovny pro digitální humanitní vědy
DNNT	díla nedostupná na trhu
DNNTO	díla nedostupná na trhu přístupná on-line po přihlášení
DNNTT	díla nedostupná na trhu přístupná přes terminál
DOI	Digital Object Identifier; digitální identifikátor objektu
DTA	Deutsches Textarchiv
FAIR	Findable, Accessible, Interoperable, Reusable; dohledatelnost, přístupnost, interoperabilita, znovuvyužitelnost
FOXML	Fedora Object XML; XML pro popis modelu digitálních objektů Fedory
GitHub	označení úložiště (často volného) zdrojového kódu, https://github.com

hOCR	standard pro zachycení dat rozpoznaných pomocí OCR
HTML	HyperText Markup Language
HTR	Handwritten Text Recognition, rozpoznávání ručně psaného textu
IIIF	International Image Interoperability Framework
IMCOMP	Image Comparator; software pro srovnání obrázků
JPEG	přípona obrazových komprimovaných souborů
JSON	JavaScript Object Notation; javascriptový objektový zápis
KNAV	Knihovna Akademie věd
LD	Levenshtein distance; Levenštejnova vzdálenost
LTP	(systém) Long-term Preservation; systém pro dlouhodobou konzervaci
MARC	MAchine-Readable Cataloging; strojově čitelnou katalogizace
MODS	Metadata Object Description Schema; schéma pro popis metadat o objektu
MZK	Moravská zemská knihovna v Brně
NDK	Národní digitální knihovna
NK	Národní knihovna České republiky
OCR	Optical Character Recognition; optické rozpoznávání znaků
ODD	One Document Does it All; formální popis standardu TEI P5
OOA-S	Ochranná organizace autorská – sdružení autorů děl výtvarného umění, architektury a obrazové složky audiovizuálních děl
OpenAPI	specifikace pro popis rozhraní REST API
PAGE XML	Page Analysis and Ground Truth Elements; kolekce standardů pro popis prvků na stránce pomocí XML
PDF	Portable Document Format, přenosný formát dokumentů
PERO	projekt vyvíjející technologie a nástroje, které zlepší a rozšíří přístupnost digitalizovaných historických dokumentů

PID	Persistent Identifier; stálý identifikátor
REST	Representational State Transfer; reprezentační přenos stavu
SaaS	Software as a Service, software jako služba
TCP	Text Creation Partnership; http://www.textcreationpartnership.org
TEI	Text Encoding Initiative, organizace pro vývoj a správu standardů pro reprezentaci textů v digitální formě
TSV	Tab-Separated Values; hodnoty oddělené tabulátorem
URI	Uniform Resource Identifier; jednotný identifikátor zdroje
UUID	Universally Unique Identifier; univerzální unikátní identifikátor
VISE	VGG Image Search Engine
VISK	Veřejná informační služba knihoven
XSD	XML Schema Definition; schéma popisující strukturu dokumentu XML
XML	Extensible Markup Language, rozšiřitelný značkovací jazyk
XQuery	XML Query; dotazovací a funkcionální programovací jazyk

11 Seznam literatury

BUENO DE LA FUENTE, Gema. bez data. What is Open Science? Introduction. In: *FOSTER* [online]. [cit. 2022-06-25]. Dostupné z: https://www.fosteropenscience.eu/content/what-open-science-introduction

ČECH, Radek. 2016. *Tematická koncentrace textu v češtině*. Prague: Ústav formální a aplikované lingvistiky. Studies in computational and theoretical linguistics. ISBN 978-80-88132-00-4.

Česká akademická federace identit eduID.cz [online]. 2021. Praha: Cesnet [cit. 2022-06-25]. Dostupné z: https://www.eduid.cz

Česká digitální knihovna: Národní agregátor digitálních knihoven [online], 2022. Praha: Knihovna AV ČR [cit. 2022-06-25]. Dostupné z: https://www.czechdigitallibrary.cz

Deutsches Textarchiv [online], 2007–2022. Berlin: Berlin-Brandenburgische Akademie der Wissenschaften [cit. 2022-06-25]. Dostupné z: https://www.deutschestextarchiv.de

Digitální knihovna [online], bez data. Brno: Moravská zemská knihovna v Brně [cit. 2022-06-25]. Dostupné z: https://www.digitalniknihovna.cz

Data (informatika). 2001. In: *Wikipedia: the free encyclopedia* [online]. San Francisco (CA): Wikimedia Foundation [cit. 2022-06-25]. Dostupné z: https://cs.wikipedia.org/wiki/Data_(informatika)

GO FAIR [online]. 2022. Hamburg – Leiden – Paris: GO FAIR International Support and Coordination Office [cit. 2022-04-13]. Dostupné z: https://www.go-fair.org

JANNIDIS, Fotis, Hubertus KOHLE a Malte REHBEIN, eds. 2017. *Digital Humanities* [online]. Stuttgart: J.B. Metzler [cit. 2022-04-13]. DOI: 10.1007/978-3-476-05446-3. ISBN 978-3-476-02622-4. Dostupné z: https://link.springer.com/book/10.1007/978-3-476-05446-3

KAPLAN, Frédéric. 2015. A Map for Big Data Research in Digital Humanities. *Frontiers in Digital Humanities* [online]. **2**: 1–7 [cit. 2022-06-25]. DOI: 10.3389/fdigh.2015.00001. ISSN 2297-2668. Dostupné z: https://doi.org/10.3389/fdigh.2015.00001

Kramerius [online]. bez data. Praha: Knihovna Akademie věd ČR [cit. 2022-06-25]. Dostupné z: https://system-kramerius.cz

KUČEROVÁ, Helena. 2019. Bibliografická metadata v sémantickém webu. *Knihovna: knihovnická revue* [online]. **30**(2): 5–35 [cit. 2022-05-30]. ISSN 1801-3252. Dostupné z: https://knihovnarevue.nkp.cz/dokumenty/2019-2/kucerova.pdf

LHOTÁK, Martin, 2020. DL4DH – Digital Libraries for Digital Humanities: nový projekt na vytěžování obsahu digitálních knihoven. *IT lib: Informačné technológie a knižnice* [online]. **2020**(4), 26–31 [cit. 2022-06-26]. ISSN 1335-793X. Dostupné z: https://itlib.cvtisr.sk/wp-content/uploads/2021/02/Lhotak.pdf

LUHMANN, Jan a Manuel BURGHARDT. 2022. Digital humanities—A discipline in its own right? An analysis of the role and position of digital humanities in the academic landscape. *Journal of the Association for Information Science and Technology*. **73**(2): 148–171. DOI: 10.1002/asi.24533. ISSN 2330-1643. Dostupné z: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.24533

MELICHAR, Marek a Jan HUTAŘ. 2013. České paměťové instituce a digitální data – historický exkurz, současný stav a předpokládaný vývoj I. *Duha: Informace o knihách a knihovnách z Moravy* [online]. **27**(4) [cit. 2022-05-21]. ISSN 1804-4255. Dostupné z: http://duha.mzk.cz/clanky/ceske-pametove-instituce-digitalni-data-historicky-exkurz--soucasny-stav-predpokladany-vyvoj

Metadata. 2011. *Databáze Národní knihovny ČR* [online]. Praha: Národní knihovna ČR [cit. 2022-06-25]. Dostupné z: https://aleph.nkp.cz/F/?func=direct&doc_number=000000543&local_base=KTD

MORSELLI, Francesca, Hella HOLLANDER, Frank UITERWAAL et al. 2020. *ZÁSADY zajištení FAIRové správy a využitelnosti dat*. Zenodo. DOI: 10.5281/zenodo.3946100. Dostupné z: http://dx.doi.org/10.5281/zenodo.3946100

Národní digitální knihovna: Digitální knihovna Kramerius [online], bez data. Praha: Národní knihovna ČR [cit. 2022-06-25]. Dostupné z: https://www.ndk.cz

NATIONAL ACADEMIES OF SCIENCES, ENGINEERING, and Medicine. 2019. *Reproducibility and Replicability in Science*. Washington D.C: National Academies Press, 257 s. ISBN 978-0-309-48619-4.

PEELS, Rik. 2019. Replicability and replication in the humanities. *Research Integrity and Peer Review* [online]. **4**(1): 1–12 [cit. 2022-06-25]. DOI: 10.1186/s41073-018-0060-4. ISSN 2058-8615. Dostupné z: https://researchintegrityjournal.biomedcentral.com/articles/10.1186/s41073-018-0060-4

PENDERS, HOLBROOK a DE RIJCKE. 2019. Rinse and Repeat: Understanding the Value of Replication across Different Ways of Knowing. *Publications* [online]. **7**(3): 1–15 [cit. 2022-06-25]. DOI: 10.3390/publications7030052. ISSN 2304-6775. Dostupné z: https://doi.org/10.3390/publications7030052

PRICE, Derek J. de Solla. 1963. *Little Science, Big Science*. New York; London: Columbia University Press, 136 s. ISBN 978-0-231-91844-2.

Registr digitalizace: Evidence dokumentů digitalizovaných v ČR [online], 2017. Praha: Národní knihovna ČR – Knihovna Akademie věd ČR – INCAD [cit. 2022-06-25]. Dostupné z: https://www.registrdigitalizace.cz

RICHTER, Vít. 2020. Zpřístupnění plných textů digitalizovaných knih a periodik prostřednictvím Národní digitální knihovny (recenzovaný článek). *Informace – zpravodaj Knihovny AV ČR* [online]. (2) [cit. 2022-04-28]. ISSN 1805-2800. Dostupné z: https://www.lib.cas.cz/casopis_informace/zpistupneni-digi-ndk/

Standardy digitalizace, 2018. In: *Národní digitální knihovna* [online]. Praha: Národní knihovna ČR [cit. 2022-06-25]. Dostupné z: https://standardy.ndk.cz/ndk/standardy--digitalizace/

Studijní on-line bible [online]. Libor Diviš [cit. 2022-06-25]. Dostupné z: https://obohu.cz/bible/

TEI: P5 Guidelines. 2022. *The Text Encoding Initiative Consortium* [online]. [cit. 2022-06-25]. Dostupné z: http://www.tei-c.org/Guidelines/P5/

Text Creation Partnership [online]. n.d. Ann Arbor: Text Creation Partnership [cit. 2022-06-25]. Dostupné z: http://www.textcreationpartnership.org

The JSON data interchange syntax. 2017. 2nd edition. Geneva: Ecma International. Dostupné z: https://www.ecma-international.org/wp-content/uploads/ECMA-404_2nd_edition_december_2017.pdf

Unicode: The World Standard for Text and Emoji [online], 2021. Mountain View: The Unicode Consortium [cit. 2022-06-25]. Dostupné z: https://home.unicode.org

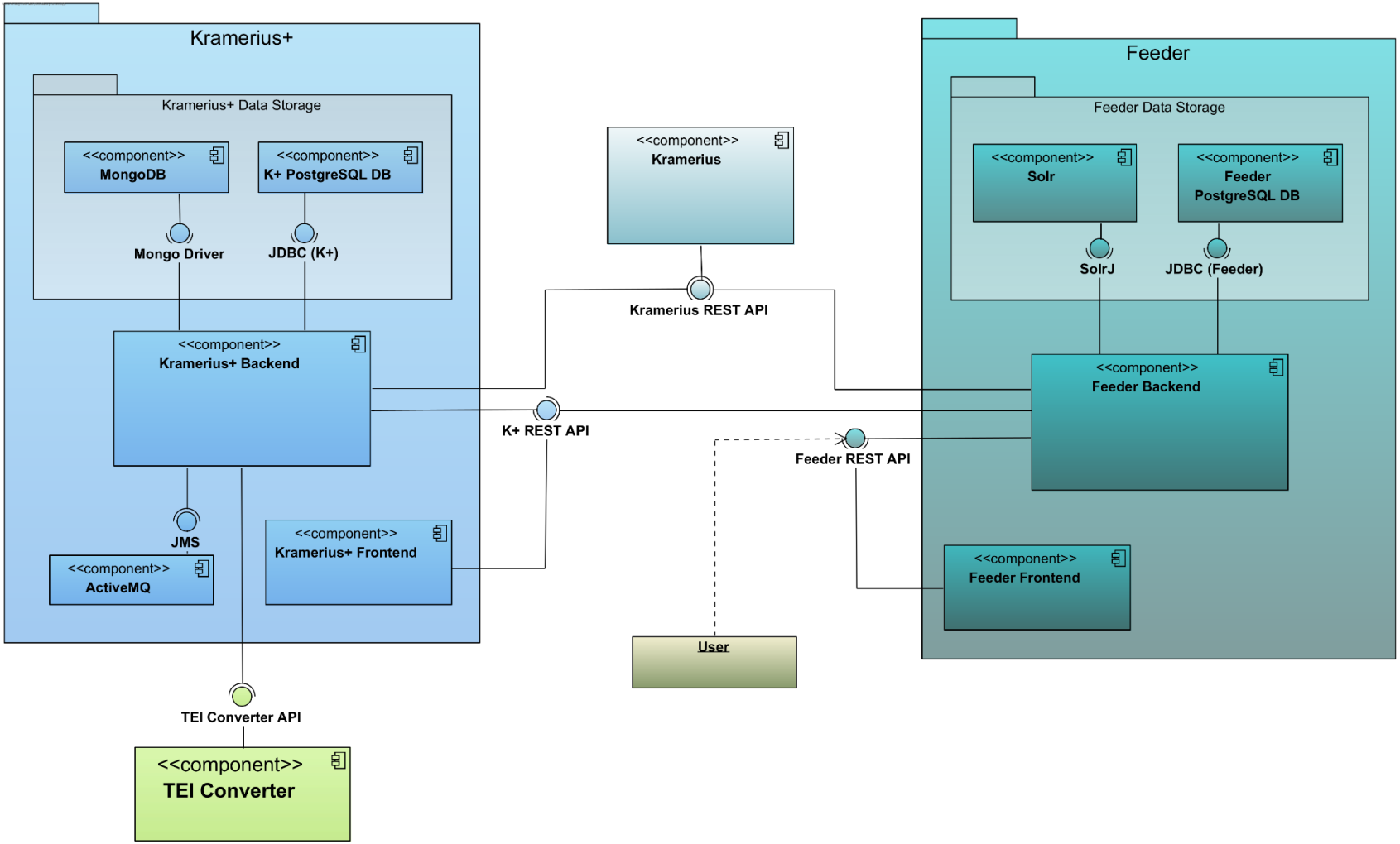
WILKENS, Matthew a Guangchen RUAN. 2020. *Geographic Locations in English-Language Literature, 1701-2011 (1.0). [Dataset]* [online]. 2020. HathiTrust Research Center [cit. 2022-06-25]. Dostupné z: https://doi.org/10.13012/2K5C-RF13

WILKINSON, Mark D., Michel DUMONTIER, IJsbrand Jan AALBERSBERG et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*. **3**(1). DOI: 10.1038/sdata.2016.18. ISSN 2052-4463. Dostupné z: https://www.nature.com/articles/sdata201618.pdf

XML Technology. 2015. *World Wide Web Consortium (W3C)* [online]. World Wide Web Consortium (W3C) [cit. 2022-06-25]. Dostupné z: https://www.w3.org/standards/xml/

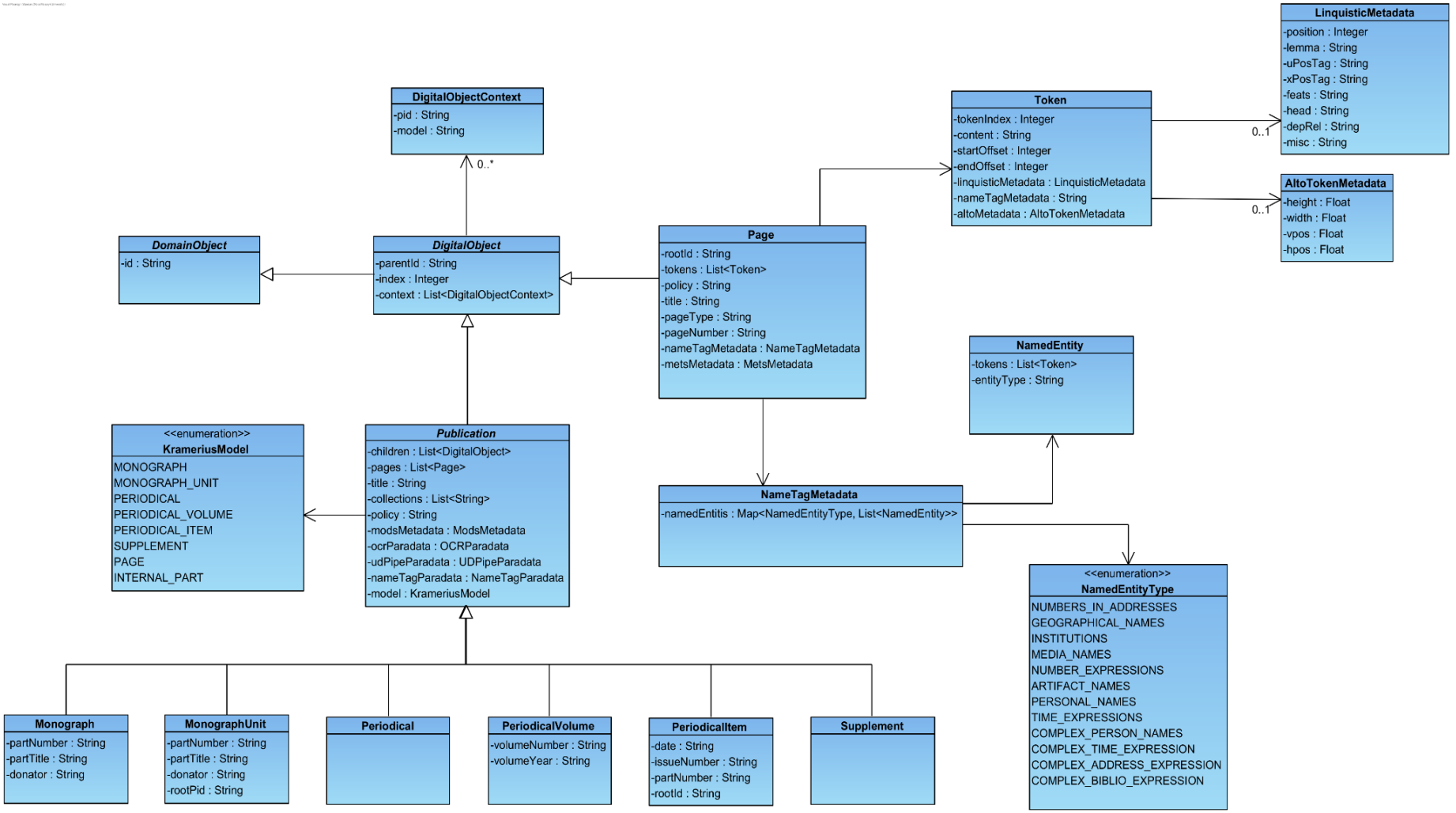
12 Přílohy

12.1 Komponenty systému DL4DH



Obrázek č. 9 Komponenty systému DL4DH

12.2 Datový model Krameria+



Obrázek č. 10 Datový model Krameria+

12.3 Ukázka exportovaných dat ve formátu prostého textu

daných podmínek používal, ovšem vždy na pravém místě [zbytek textu vynechán]

12.4 Ukázka exportovaných dat ve formátu ALTO

```
<Layout>
  <Page ID="Page1" PHYSICAL_IMG_NR="1">
   <PrintSpace HEIGHT="2482" WIDTH="1576" VPOS="0" HPOS="0">
    <TextBlock ID="BlockId-DDA443A6-B713-4823-B950-6CE779FA18C9-"
       HEIGHT="1712" WIDTH="935" VPOS="356" HPOS="261"
       STYLEREFS="StyleId-2A800981-7356-45BE-9520-36D40D439C4E- font1">
     <TextLine HEIGHT="50" WIDTH="912" VPOS="361" HPOS="279">
      <String CONTENT="daných" HEIGHT="48" WIDTH="154" VPOS="361" HPOS="279"/>
      <SP WIDTH="26" VPOS="361" HPOS="434"/>
      <String CONTENT="podmínek" HEIGHT="48" WIDTH="211" VPOS="362" HPOS="461"/>
      <SP WIDTH="32" VPOS="362" HPOS="673"/>
      <String CONTENT="používal," HEIGHT="47" WIDTH="192" VPOS="363" HPOS="706"/>
      <SP WIDTH="33" VPOS="376" HPOS="899"/>
      <String CONTENT="ovšem" HEIGHT="35" WIDTH="131" VPOS="365" HPOS="933"/>
      <SP WIDTH="26" VPOS="376" HPOS="1065"/>
      <String CONTENT="vždy" HEIGHT="48" WIDTH="99" VPOS="363" HPOS="1092"/>
     </TextLine>
     <TextLine HEIGHT="48" WIDTH="909" VPOS="425" HPOS="280">
      <String CONTENT="na" HEIGHT="23" WIDTH="52" VPOS="437" HPOS="280"/>
      <SP WIDTH="26" VPOS="437" HPOS="333"/>
```

```
      <String CONTENT="pravém" HEIGHT="47" WIDTH="160" VPOS="425" HPOS="360"/>
      <SP WIDTH="37" VPOS="438" HPOS="521"/>
      <String CONTENT="místě" HEIGHT="35" WIDTH="110" VPOS="426" HPOS="559"/>
      <!-- [zbytek textu vynechán] -->
     </TextLine>
    </TextBlock>
   </PrintSpace>
  </Page>
 </Layout>
```

12.5 Ukázka exportovaných dat ve formátu TSV

```
page_id  token lemma position nameTag  udpipe.feats.Abbr  udpipe.feats.AdpType  udpipe.feats.Animacy  udpipe.feats.Aspect
   udpipe.feats.Case  udpipe.feats.Degree   udpipe.feats.Gender   udpipe.feats.Gender[psor]   udpipe.feats.Mood
   udpipe.feats.NameType udpipe.feats.NumForm  udpipe.feats.NumType  udpipe.feats.Number   udpipe.feats.Number[psor]
   udpipe.feats.Person   udpipe.feats.Polarity udpipe.feats.Poss  udpipe.feats.PrepCase udpipe.feats.PronType udpipe.feats.Reflex
   udpipe.feats.Tense udpipe.feats.Variant  udpipe.feats.VerbForm udpipe.feats.Voice udpipe.misc.SpaceAfter
   udpipe.misc.TokenRange
uuid:796cde10-2d7c-11e4-a8ab-001018b5eb5c  danýchdaný  1                GenPosFem               Plur        Pos
                        0:6
uuid:796cde10-2d7c-11e4-a8ab-001018b5eb5c  podmínek podmínka 2                Gen   Fem               Plur
   Pos                           7:15
uuid:796cde10-2d7c-11e4-a8ab-001018b5eb5c  používal používat 3             Imp      Masc                Sing
   Pos            Past     Part  ActNo 16:24
uuid:796cde10-2d7c-11e4-a8ab-001018b5eb5c  ,  ,  4        24:25
uuid:796cde10-2d7c-11e4-a8ab-001018b5eb5c  ovšem ovšem 5        26:31
```

```
uuid:796cde10-2d7c-11e4-a8ab-001018b5eb5c  vždy  vždy  6
   Tot                   32:36
uuid:796cde10-2d7c-11e4-a8ab-001018b5eb5c  na na 7        Prep        Loc
                  37:39
uuid:796cde10-2d7c-11e4-a8ab-001018b5eb5c  pravémpravý 8                 LocPosNeut                  Sing           Pos
                           40:46
uuid:796cde10-2d7c-11e4-a8ab-001018b5eb5c  místě místo 9                 Loc   Neut                  Sing           Pos
                           47:52
```

12.6 Ukázka exportovaných dat ve formátu CSV

```
page_id,token,lemma,position,nameTag,udpipe.feats.Abbr,udpipe.feats.AdpType,udpipe.feats.Animacy,udpipe.feats.Aspect,udpipe.fea
ts.Case,udpipe.feats.Degree,udpipe.feats.Gender,udpipe.feats.Gender[psor],udpipe.feats.Mood,udpipe.feats.NameType,udpipe.feats.
NumForm,udpipe.feats.NumType,udpipe.feats.Number,udpipe.feats.Number[psor],udpipe.feats.Person,udpipe.feats.Polarity,udpipe.fea
ts.Poss,udpipe.feats.PrepCase,udpipe.feats.PronType,udpipe.feats.Reflex,udpipe.feats.Tense,udpipe.feats.Variant,udpipe.feats.Ve
rbForm,udpipe.feats.Voice,udpipe.misc.SpaceAfter,udpipe.misc.TokenRange
uuid:796cde10-2d7c-11e4-a8ab-001018b5eb5c,daných,daný,1,,,,,,Gen,Pos,Fem,,,,,Plur,,,Pos,,,,,,,,,0:6
uuid:796cde10-2d7c-11e4-a8ab-001018b5eb5c,podmínek,podmínka,2,,,,,,Gen,,Fem,,,,,Plur,,,Pos,,,,,,,,,7:15
uuid:796cde10-2d7c-11e4-a8ab-001018b5eb5c,používal,používat,3,,,,,Imp,,,Masc,,,,,Sing,,,Pos,,,,,Past,,Part,Act,No,16:24
uuid:796cde10-2d7c-11e4-a8ab-001018b5eb5c,",",",",",4,,,24:25
uuid:796cde10-2d7c-11e4-a8ab-001018b5eb5c,ovšem,ovšem,5,,,26:31
uuid:796cde10-2d7c-11e4-a8ab-001018b5eb5c,vždy,vždy,6,,,,,,,,,,,,,,,,,,Tot,,,,,,,32:36
uuid:796cde10-2d7c-11e4-a8ab-001018b5eb5c,na,na,7,,,Prep,,,Loc,,,,,,,,,,,,,,,,,,,37:39
uuid:796cde10-2d7c-11e4-a8ab-001018b5eb5c,pravém,pravý,8,,,,,Loc,Pos,Neut,,,,,Sing,,,Pos,,,,,,,,,40:46
uuid:796cde10-2d7c-11e4-a8ab-001018b5eb5c,místě,místo,9,,,,,Loc,,Neut,,,,,Sing,,,Pos,,,,,,,,,47:52
```

12.7 Ukázka exportovaných dat ve formátu TEI

```
<pb xml:id="uuid-796cde10-2d7c-11e4-a8ab-001018b5eb5c" n="11"/>
 <p>
  <s>
   <w n="1" pos="ADJ" msd="Case=Gen|Degree=Pos|Gender=Fem|Number=Plur|Polarity=Pos" lemma="daný">daných</w>
   <w n="2" pos="NOUN" msd="Case=Gen|Gender=Fem|Number=Plur|Polarity=Pos" lemma="podmínka">podmínek</w>
   <w n="3" pos="VERB" msd="Aspect=Imp|Gender=Masc|Number=Sing|Polarity=Pos|Tense=Past|VerbForm=Part|Voice=Act"
lemma="používat">používal</w>
   <pc n="4" pos="PUNCT" join="left" msd="" lemma=",">,</pc>
   <w n="5" pos="CCONJ" msd="" lemma="ovšem">ovšem</w>
   <w n="6" pos="ADV" msd="PronType=Tot" lemma="vždy">vždy</w>
   <w n="7" pos="ADP" msd="AdpType=Prep|Case=Loc" lemma="na">na</w>
   <w n="8" pos="ADJ" msd="Case=Loc|Degree=Pos|Gender=Neut|Number=Sing|Polarity=Pos" lemma="pravý">pravém</w>
   <w n="9" pos="NOUN" msd="Case=Loc|Gender=Neut|Number=Sing|Polarity=Pos" lemma="místo">místě</w>
   <!-- [zbytek textu vynechán] -->
  </s>
  <!-- [zbytek textu vynechán] -->
 </p>
```

12.8 Ukázka exportovaných dat ve formátu JSON


```
{
   "model": "page",
   "id": "uuid:796cde10-2d7c-11e4-a8ab-001018b5eb5c",
   "created": "2022-05-15T23:49:55.109Z",
   "index": 14,
   "context": [],
   "tokens": [
    {
     "tokenIndex": 0,
     "content": "daných",
     "startOffset": 0,
     "endOffset": 6,
     "linguisticMetadata": {
      "position": 1,
      "lemma": "daný",
      "feats": "Case=Gen|Degree=Pos|Gender=Fem|Number=Plur|Polarity=Pos",
      "head": "2",
      "depRel": "amod",
      "misc": "TokenRange=0:6",
      "uPosTag": "ADJ",
      "xPosTag": "AAFP2----1A----"
     }
    },
    {
     "tokenIndex": 1,
     "content": "podmínek",
     "startOffset": 7,
     "endOffset": 15,
     "linguisticMetadata": {
      "position": 2,
      "lemma": "podmínka",
      "feats": "Case=Gen|Gender=Fem|Number=Plur|Polarity=Pos",
      "head": "3",
      "depRel": "obl:arg",
      "misc": "TokenRange=7:15",
      "uPosTag": "NOUN",
      "xPosTag": "NNFP2-----A----"
```

```
      }
     },
     {
      "tokenIndex": 2,
      "content": "používal",
      "startOffset": 16,
      "endOffset": 24,
      "linguisticMetadata": {
       "position": 3,
       "lemma": "používat",
       "feats":
"Aspect=Imp|Gender=Masc|Number=Sing|Polarity=Pos|Tense=Past|VerbForm=Part|Voice=Ac
t",
       "head": "0",
       "depRel": "root",
       "misc": "SpaceAfter=No|TokenRange=16:24",
       "uPosTag": "VERB",
       "xPosTag": "VpYS---XR-AA---"
      }
     },
     {
      "tokenIndex": 3,
      "content": ",",
      "startOffset": 24,
      "endOffset": 25,
      "linguisticMetadata": {
       "position": 4,
       "lemma": ",",
       "head": "6",
       "depRel": "punct",
       "misc": "TokenRange=24:25",
       "uPosTag": "PUNCT",
       "xPosTag": "Z:-------------"
      }
     },
     {
      "tokenIndex": 4,
      "content": "ovšem",
      "startOffset": 26,
      "endOffset": 31,
```

```
 "linguisticMetadata": {
  "position": 5,
  "lemma": "ovšem",
  "head": "6",
  "depRel": "cc",
  "misc": "TokenRange=26:31",
  "uPosTag": "CCONJ",
  "xPosTag": "J^-------------"
 }
},
{
 "tokenIndex": 5,
 "content": "vždy",
 "startOffset": 32,
 "endOffset": 36,
 "linguisticMetadata": {
  "position": 6,
  "lemma": "vždy",
  "feats": "PronType=Tot",
  "head": "3",
  "depRel": "conj",
  "misc": "TokenRange=32:36",
  "uPosTag": "ADV",
  "xPosTag": "Db-------------"
 }
},
{
 "tokenIndex": 6,
 "content": "na",
 "startOffset": 37,
 "endOffset": 39,
 "linguisticMetadata": {
  "position": 7,
  "lemma": "na",
  "feats": "AdpType=Prep|Case=Loc",
  "head": "9",
  "depRel": "case",
  "misc": "TokenRange=37:39",
  "uPosTag": "ADP",
  "xPosTag": "RR--6----------"
```

```
      }
     },
     {
      "tokenIndex": 7,
      "content": "pravém",
      "startOffset": 40,
      "endOffset": 46,
      "linguisticMetadata": {
       "position": 8,
       "lemma": "pravý",
       "feats": "Case=Loc|Degree=Pos|Gender=Neut|Number=Sing|Polarity=Pos",
       "head": "9",
       "depRel": "amod",
       "misc": "TokenRange=40:46",
       "uPosTag": "ADJ",
       "xPosTag": "AANS6----1A----"
      }
     },
     {
      "tokenIndex": 8,
      "content": "místě",
      "startOffset": 47,
      "endOffset": 52,
      "linguisticMetadata": {
       "position": 9,
       "lemma": "místo",
       "feats": "Case=Loc|Gender=Neut|Number=Sing|Polarity=Pos",
       "head": "6",
       "depRel": "orphan",
       "misc": "TokenRange=47:52",
       "uPosTag": "NOUN",
       "xPosTag": "NNNS6-----A----"
      }
     }
    …
}
```