ČASOPIS PRO MODERNÍ FILOLOGII 105, 2023, Č. 2, S. 274–292

Blýskání na lepší data z českých digitálních knihoven[1]

Boris Lehečka (Brno)

THERE ARE BETTER DATA AHEAD FROM CZECH DIGITAL LIBRARIES
In the humanities, analysis of primary and secondary literature is an important area of research work. Besides language corpora, digital libraries, which digitized approximately 98.7 million pages in the Czech Republic between 1992 and 2022, can be considered a suitable source of written texts in recent years. The article presents an example from abroad and gives a brief overview of data sources in the Czech environment. It focuses on the recently completed DL4DH project, which aims to offer researchers access to large volumes of data from the Kramerius digital library in standardized formats (plain text, ALTO, CSV/TSV, TEI, JSON) not only through a new web application but also through a REST API. To make the subsequent analysis of the publications as easy as possible, the downloaded data can include enrichment data from the UDPipe and NameTag tools developed and operated by the LINDAT/CLARIAH-CZ research infrastructure.

KLÍČOVÁ SLOVA
velká data, digitální knihovna, digitální humanitní vědy, výzkumná infrastruktura, autorský zákon

KEYWORDS
big data, digital library, digital humanities, research infrastructure, copyright law

DOI
https://doi.org/10.14712/23366591.2023.2.7

1. ÚVOD

Poslední dobou se objevuje stále více pozoruhodných výstupů aplikovaného výzkumu, které vzbuzují oprávněný zájem ze strany laické i odborné veřejnosti. Namátkou můžeme zmínit např. systém DALL·E 2 (2022),[2] který na základě slovního popisu generuje realisticky vypadající výtvarná díla (obrazy); Charles Translator for Ukraine (2022)[3] určený pro strojový překlad mezi češtinou a ukrajinštinou v reakci na ruskou agresi na Ukrajině a na příliv ukrajinských běženců do Česka; systém ChatGPT (2022),[4] který písemně komunikuje s uživatelem: odpovědi na kladené otázky

1 Tato studie vznikla v rámci Dlouhodobé koncepce pro rozvoj výzkumné organizace — Moravská zemská knihovna v Brně.
2 Viz též <https://openai.com/dall-e-2/>.
3 Viz též <https://www.mff.cuni.cz/cs/verejnost/aktuality/vedci-z-matfyzu-vyvinuli-automaticky-prekladac-mezi-cestinou-a-ukrajinstinou>.
4 Viz též <https://openai.com/blog/chatgpt/>.

vypadají věcně správně, názorně, edukativně, přičemž berou v potaz fakta z předchozí konverzace. Takový posun v oblasti strojového učení a umělé inteligence by nebyl možný bez nových technologií, nových výpočetních metod a zejména bez velkých objemů dat, která jsou pro trénování těchto technologií nezbytná.[5]

Podívejme se, jaké možnosti nabízí badatelům v tuto chvíli české prostředí: s jakými daty můžou badatelé pracovat a na základě jakých autorským zákonem daných pravidel. Uvedený přehled dostupných zdrojů a nástrojů představuje pouze výběr z mnoha možností a zaměřuje se specificky na digitální knihovny; kompendium rovněž nemusí sloužit pouze pro budování systémů s umělou inteligencí, ale najde uplatnění při jakémkoli výzkumu v oblasti humanitních a sociálních věd.[6]

2. DATA

Z hlediska podoby dat pro výzkum můžeme rozlišovat mezi daty surovými a zpracovanými. Aby bylo možné údaje využít pro konkrétní badatelský záměr, musí se v drtivé většině případů nějakým způsobem zpracovat. Způsob a míra zpracování záleží na výzkumné otázce, na niž má analýza odpovědět. Za surová data může považovat například obrázky a texty z českých digitálních knihoven. Ke strukturovaným datům se badatelé dostanou ve specializovaných repozitářích. Například v úložišti výzkumné infrastruktury LINDAT/CLARIAH-CZ najdeme zejména jazykově zpracovaná data,[7] výzkumná infrastruktura Český sociálněvědní datový archiv (ČSDA) naproti tomu shromažďuje data z českých sociálněvědních šetření, výzkumů veřejného mínění a mezinárodních šetření s českou účastí.[8]

Přístup k datům dostupným na internetu bývá primárně navržen tak, aby k nim uživatel přistupoval prostřednictvím uživatelského rozhraní, obvykle webové aplikace. Další způsob představují webové aplikace navržené tak, aby se k datům mohlo přistupovat automatizovaně pomocí programových prostředků. V dřívějších dobách k tomu sloužily různé protokoly a technologie, např. SOAP,[9] v poslední době je populární programové rozhraní v kombinaci s architekturou REST, tzv. REST API.[10]

5 Větší objemy dat mají vliv nejen na větší relevantnost odpovědi, kterou systém vrací, ale také na to, že systémy mohou adekvátně reagovat na širší repertoár vstupů, které od uživatelů přicházejí. Nejnovější technologie si dokážou poradit se stovkami miliard slov a miliony parametrů.

6 Autor děkuje oběma anonymním recenzentům za jejich cenné připomínky a náměty.

7 Viz <https://lindat.mff.cuni.cz/repository/> a <https://bit.ly/cmf-lindat-riv-2020>.

8 Viz <https://www.soc.cas.cz/oddeleni/cesky-socialnevedni-datovy-archiv> a <https://bit.ly/cmf-csda-riv-2020>.

9 Data spolu s relevantními metadaty se přenášejí ve formátu XML; o standard se stará konsorcium W3C, viz jednotlivé součásti <https://www.w3.org/TR/soap12-part0/>, <https://www.w3.org/TR/soap12-part1/> a <https://www.w3.org/TR/soap12-part2/>.

10 K přenosu a úpravě dat se využívají standardní součásti protokolu HTTP; data se přenášejí obvykle ve formátu XML nebo JSON; dokumentaci programového rozhraní a standardizaci v tomto případě zaručuje specifikace OpenAPI (<https://swagger.io/specification/>).

3. ZAHRANIČNÍ ZKUŠENOSTI

V zahraničí mají repozitáře zaměřené na poskytování velkých objemů dat své místo již dlouhou řadu let. Jedním z nich je projekt HathiTrust (2008–2023), který sdružuje zhruba 150 zahraničních knihoven, zejména z anglojazyčných zemí. Průběžně digitalizované publikace se stávají součástí jedné digitální knihovny, pojmenované HathiTrust Digital Library. V roce 2022 obsahovala přibližně 17,1 mil. publikací, z toho 6,57 mil. položek tvořila volná díla[11] a 10,55 mil. podléhala ochraně majetkových práv podle aktuálně platných autorských zákonů.[12] 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, případně je stáhnout ve formátu PDF s textovou vrstvou, např. prostřednictvím Google Books (v tomto případě jsou však obrázky černobílé, nikoli plně barevné).

Badatelé mohou pracovat s kolekcemi vybraných dokumentů, ale kvůli autorským právům mají přístup pouze k agregovaným datům z těchto děl (např. k frekvenci slovních tvarů na jednotlivých stranách). Uživatelé z partnerských institucí projektu mají v rámci HathiTrust Research Center k dispozici nástroje pro další analýzu[13] zvolené sady publikací. Jedná se o rozpoznání pojmenovaných entit,[14] tzv. „extrahované prvky“,[15] identifikace témat metodou InPhO,[16] počty tokenů (včetně morfologické analýzy) a slovní mraky.[17] 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 dále k dispozici programové knihovny v jazyce Python pro práci s vytvořenými kolekcemi metadat. Pracovníci z partnerských institucí mohou žádat o vytvoření samostatného virtuálního počítače s nezbytným softwarem a přístupem ke všem textovým zdrojům HathiTrustu, které mohou pro svůj výzkum kombinovat s vlastními daty a metadaty, případně softwarem.

11 Viz <https://cs.wikipedia.org/wiki/Volné_dílo>.

12 Pouhá dvě promile z tohoto objemu (přibližně 40 tis. publikací) tvoří dokumenty v češtině.

13 Viz <https://analytics.hathitrust.org>, zdrojové kódy jsou dostupné na <https://github.com/htrc>.

14 Viz <https://analytics.hathitrust.org/algorithms/Named_Entity_Recognizer>.

15 Viz <https://wiki.htrc.illinois.edu/display/COM/HTRC+Derived+Datasets>.

16 Viz <https://inpho.github.io/topic-explorer/>.

17 Viz <https://analytics.hathitrust.org/algorithms/Token_Count_and_Tag_Cloud_Creator>.

4. ČESKÉ PROSTŘEDÍ

V českém prostředí mají k projektu HathiTrust velmi blízko digitální knihovny známé pod označením Kramerius.[18] Než se této platformě, která je primárně určena pro digitalizované knihovní sbírky, monografie a periodika, budeme věnovat podrobněji, zmíníme se o dalších specializovaných zdrojích, které se dají využít pro badatelský výzkum zaměřený na české prostředí (jazykově i teritoriálně).

4.1 MANUSCRIPTORIUM

Digitální knihovna Manuscriptorium (2023) umožňuje snadný přístup k soustředěným informacím o historických fondech ze stovek paměťových institucí. Jedná se o celosvětově největší digitální knihovnu zaměřenou na starší písemné dědictví, která obsahuje přes 154 tis. digitálních dokumentů různého typu (rukopisy, inkunábule, staré tisky, 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. Virtuální prostředí Manuscriptoria např. zajišťuje práci s dokumenty, jež jsou kompatibilní s technologiemi IIIF,[19] z libovolného externího zdroje, a to díky integraci prohlížeče Mirador.[20] Data každého digitalizátu jsou popsána dle schéma TEI P5 ENRICH,[21] v rámci digitální knihovny má digitalizát přidělený identifikátor URI a jeho použití, popř. použití jeho metadat, 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.

4.2 MONASTERIUM

Projekt Monasterium (2023) se věnuje digitalizovaným listinám, zajišťuje přístup k digitálním dokumentům (přes 500 tis. listin) ze 176 evropských archivů,[22] z toho 11 českých. Pro metadata o jednotlivých listinách se používá specializovaný formát XML CEI.[23] Badatelé mohou stahovat obrázky dokumentů ve formátu JPEG nebo PDF (ve vysokém rozlišení). Na stránkách projektu je také informace, že „[v]ešker á práva na zveřejňování a rozšiřování reprodukcí listin v obrazové formě náleží jednotlivým vlastníkům archiválií".[24] Ne vždy je však z jednotlivých webových stránek s digita-

18 Kramerius je původní označení opensourcové platformy, která slouží k publikování digitálních publikací.

19 Viz <https://iiif.io>.

20 Viz <https://projectmirador.org>.

21 Viz <https://www.manuscriptorium.com/cs/tei-p5-enrich-schema-cs>. Jelikož však samotné dokumenty s kořenovým elementem TEI neobsahují označení jmenného prostoru, neodpovídají dokumenty z Manuscriptoria deklarovanému standardu.

22 Seznam institucí je dostupný zde: <https://www.monasterium.net/mom/fonds>.

23 Charters Encoding Initiative, viz <https://www.cei.lmu.de>.

24 Viz <https://www.monasterium.net/mom/terms-of-use>.

lizáty nebo z jejich zdrojových dat ve formátu CEI patrné, jaké licenční podmínky se musí při nakládání s pramenem dodržovat, takže je jistější kontaktovat instituci, která danou listinu spravuje.

4.3 CZECH MEDIEVAL SOURCES ONLINE

Webová aplikace Czech medieval sources FONTES (2023) obsahuje skeny sekundárních zdrojů (edic) pro studium dějin českého středověku (ke konci roku 2022 se jednalo o 900 svazků o celkovém objemu téměř 261 tis. stran). Většinou jde o volná díla, u nichž již uplynulo 70 let od úmrtí autora. Badatelé mají k dispozici digitální obrázky a plnotextové prohledávání. Publikace lze procházet po jednotlivých stranách. Centrum medievistických studií spravuje rovněž aplikaci HyperFontes (2023), digitální repertorium pramenů k českému středověku, tj. záznamy a informace o autorech a dílech bohemikálního charakteru. Tyto zdroje jsou součástí výzkumné infrastruktury středověkých a raně novověkých bohemikálních pramenů MEMORIA, která se začlenila do velké výzkumné infrastruktury LINDAT/CLARIAH-CZ.

4.4 LINDAT/CLARIAH-CZ

Výzkumná infrastruktura LINDAT/CLARIAH-CZ (2023), jejíž historie sahá do roku 2010, nabízí badatelům celou řadu specializovaných i obecně zaměřených zdrojů, mj. Bibliografii dějin Českých zemí (2013) — bibliografická data, velké množství prohledávatelných korpusů[25] (např. ParCzech — česká parlamentní data zkompilovaná do korpusu s lingvistickými anotacemi) nebo tematicky zaměřené zdroje (Vokabulář webový — textové, obrazové a zvukové zdroje k poznávání historické češtiny; Arne Novák — digitální knihovna prof. Arna Nováka, literárního vědce, kritika, historika a esejisty; Malach — digitální archiv orálně historických rozhovorů ap.). Významné těžiště představuje repozitář[26] pro jazyková data a nástroje na zpracování textu. K jeho přednostem patří, že jsou uložená data i nástroje opatřeny kvalitními metadaty včetně licenčních podmínek, takže je lze snadno dohledat, citovat a používat. K datům a nástrojům s licencí, která neumožňuje bezplatné a volné sdílení, je možné přistoupit až po přihlášení k repozitáři. Lze využít přihlášení prostřednictvím institucionálního účtu u akademické instituce, která je součástí federace EduGAIN;[27] obvykle je tedy možné se přihlásit pomocí účtu v síti eduroam,[28] v případě potřeby lze požádat o individuální přístup. Přihlášení je nezbytné pro badatele, kteří chtějí svá data v repozitáři uložit.

Ke konci roku 2022 obsahoval tento repozitář 492 položek, které se týkaly českého jazyka a nesly s sebou obsah v podobě přiložených souborů.[29]

25 Viz <https://lindat.mff.cuni.cz/services/kontext/corpora/corplist>.

26 Digitální úložiště spravované Lindatem vychází z opensourcového řešení DSpace (<https://dspace.lyrasis.org>); zdrojový kód je k dispozici v úložišti GitHub (<https://github.com/ufal/clarin-dspace>).

27 Viz <https://edugain.org>.

28 Viz <https://www.eduroam.cz>.

29 Viz dotaz <https://bit.ly/lindat-ceske-zdroje>. Pro ostatní jazyky bylo k dispozici 1406 položek, celkem LINDAT/CLARIAH-CZ evidoval 2099 položek.

Z hlediska práce s daty jsou důležité také nástroje, které LINDAT/CLARIAH-CZ pro práci s jazykovými daty nabízí. Vzhledem k zaměření našeho článku upozorňujeme zejména na dva z nich: UDPipe a NameTag, které lze využívat jako samostatné aplikace nebo prostřednictvím webové služby.

Aplikace UDPipe 2 (Straka, 2016)[30] texty tokenizuje, lemmatizuje a opatřuje morfologickou a syntaktickou anotací. Morfologická anotace má podobu jak pozičních značek,[31] tak universal POS tags[32] a universal features,[33] které vycházejí z mezinárodního standardu Universal Dependencies.[34] Software samotný je k dispozici pod licencí Mozilla Public License 2.0,[35] datové modely pro více než šedesát jazyků[36] jsou dostupné pod licencí CC BY-NC-SA.[37]

Aplikace NameTag 2 (Straka, 2014)[38] slouží k identifikaci pojmenovaných entit (tj. jmen osob, institucí, geopolitických celků, časových údajů ap.). Datové modely[39] existují pro pět jazyků. Pro program a jeho datové modely platí obdobné licenční podmínky jako u předchozího nástroje.

4.5 JAZYKOVÉ KORPUSY

Na práci s velkými objemy dat se od svých počátků zaměřovaly jazykové korpusy. 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".[40] 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.).

Zpočátku vznikaly textové korpusy na základě tištěných předloh, později i na základě digitálních dokumentů a v poslední době vznikají také korpusy sestavené z textů dostupných na webových stránkách (viz např. projekt Aranea (Benko, 2014)). V současnosti existuje velké množství korpusů různých národních jazyků s obecným i specifickým zaměřením na určité oblasti jazyka (poezie, historické texty, parlamentní projevy apod.).

Budování korpusu pro český jazyk se ujal Ústav Českého národního korpusu FF UK, který mj. od roku 2000 co pět let zveřejňuje reprezentativní referenční korpus synchronní češtiny o objemu cca 100 mil. slov; poslední verze SYN2020 (2020)[41] obsa-

30 Zdrojový kód viz <https://github.com/ufal/udpipe>.
31 Viz <https://wiki.korpus.cz/doku.php/:seznamy:tagy>.
32 Viz <https://universaldependencies.org/u/pos/index.html>.
33 Viz <https://universaldependencies.org/u/feat/index.html>.
34 Viz <https://universaldependencies.org>.
35 Viz <https://www.mozilla.org/en-US/MPL/2.0/>.
36 Viz <https://ufal.mff.cuni.cz/udpipe/2/models>.
37 Viz <https://creativecommons.org/licenses/by-nc-sa/4.0/>.
38 Zdrojový kód viz <https://github.com/ufal/nametag>.
39 Viz <https://ufal.mff.cuni.cz/nametag/2/models>.
40 Viz <https://wiki.korpus.cz/doku.php/pojmy:korpus>.
41 Viz <https://wiki.korpus.cz/doku.php/cnk:syn2020>.

huje 121 826 797 pozic. Spojením výše uvedených synchronních korpusů spolu s korpusy publicistickými i dosud nezpracovanými texty vzniká korpus SYN (2022), který ve verzi 11[42] z roku 2022 obsahuje 6 067 313 960 pozic.[43]

Pro češtinu existuje také několik webových korpusů: Web Corpus of Czech (Spoustová, 2012) — z roku 2012, 628 332 859 pozic; Araneum Bohemicum Maximum (Czech, 15.04; Benko, 2015)[44] — z roku 2015, 3 198 768 569 pozic; nebo poslední Araneum Bohemicum IV Maximum (Czech, 20.03; Benko, 2020)[45] — z roku 2020, 7 103 994 584 pozic. Ke specializovaným zdrojům můžeme řadit např. Korpus českého verše (2021) (lemmatizovaný, foneticky, morfologicky, metricky a stroficky anotovaný korpus české poezie 19. a počátku 20. století; 1689 básnických sbírek; 14 592 037 slov)[46] nebo Staročeskou textovou banku (bez data) (transkribovaná nevyvážená kolekce textů staročeských literárních památek z období přibližně mezi lety 1300 až 1500; 6 953 501 pozic) ap.

K práci s korpusovými daty slouží specializované aplikace nazývané korpusové manažery. Všechny výše zmiňované zdroje využívají manažery, které vycházejí z volně dostupné aplikace NoSketch Engine.[47] Aplikace KonText[48] ke konci roku 2022 přišla s přístupem ke korpusům pomocí programového rozhraní REST API.[49] Programové rozhraní již delší dobu nabízí komerční produkt Sketch Engine.[50]

Výzkumná infrastruktura LINDAT/CLARIAH-CZ začlenila v roce 2019 mezi prezentační nástroje rovněž manažer TEITOK,[51] určený pro tzv. „živé" korpusy, které se průběžně mění a teprve v určitých fázích vývoje se z nich stávají verzované statické korpusy. TEITOK využívá pro korpusové dotazy volně dostupnou aplikaci IMS Open Corpus Workbench (CWB).[52]

42 Viz <https://wiki.korpus.cz/doku.php/cnk:syn:verze11>.

43 Kromě tohoto díla vznikají péčí ÚČNK i další, více či méně specializované korpusy, např. Totalita (korpus psaného jazyka komunistického režimu, 15 350 741 pozic), CzeSL-plain 2 (žákovský korpus češtiny nerodilých mluvčích, 2 320 678 pozic), ORAL (referenční korpus neformální mluvené češtiny, 6 361 707 pozic), DIAKORP (verzovaný korpus diachronní složky ČNK, 4 128 874 pozic) aj.

44 Dostupný pouze pro registrované uživatele.

45 Dostupný pouze pro registrované uživatele.

46 Veřejně dostupná data z repozitáře na GitHubu (<https://github.com/versotym/corpusCzechVerse>) obsahují kvůli licenčním podmínkám pouze 1305 sbírek, 66 428 básní, 2 310 917 veršů, 12 636 867 slov, 15 399 220 pozic (tj. slov a interpunkce).

47 Viz <https://nlp.fi.muni.cz/trac/noske>.

48 Viz <https://github.com/czcorpus/kontext>.

49 Viz <https://wiki.korpus.cz/doku.php/manualy:api> (základní informace) a <https://github.com/czcorpus/kontext/wiki/HTTP-API (podrobná dokumentace).

50 Viz <https://www.sketchengine.eu/documentation/api-documentation/>.

51 Viz <https://www.teitok.org>; zdrojový kód: <https://gitlab.com/maartenes/TEITOK> a <https://github.com/ufal/teitok-tools>.

52 Viz <https://cwb.sourceforge.io>; zdrojový kód: <https://sourceforge.net/p/cwb/code/HEAD/tree/>.

5. AUTORSKÝ ZÁKON

Ať už badatel pracuje s jakýmikoli daty, na něž můžeme pohlížet jako na databázi nebo na autorské dílo, měl by dodržovat základní pravidla nakládání s nimi daná autorským zákonem, tj. zákonem č. 121/2000 Sb. (dále též AZ).[53] Tato právní norma např. v § 31 stanovuje pravidla pro užití citace nebo se v § 92 vyjadřuje k vytěžování či zužitkovávání databází. Dne 5. ledna 2023 vstoupila v platnost podstatná aktualizace AZ, která vychází z adaptace směrnice Evropského parlamentu a Rady č. 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.[54] Schválené znění má mj. vliv na vytěžování textů a dat za účelem získávání nových poznatků a objevování nových trendů, jde zejména o § 39c a § 39d.

§ 39c: *Licence k rozmnožování díla pro účely automatizované analýzy textů nebo dat*

(1) Do práva autorského nezasahuje ten, kdo zhotoví rozmnoženinu díla za účelem automatizované analýzy textů nebo dat v digitální podobě, prováděné za účelem získání informací, zahrnujících mimo jiné vzory, tendence a souvztažnosti; takto zhotovenou rozmnoženinu je oprávněn uchovat pouze po dobu nezbytnou pro účely této automatizované analýzy textů nebo dat.

(2) Ustanovení odstavce 1 se nepoužije pro rozmnoženiny díla, jehož autor si užití podle odstavce 1 výslovně vyhradil vhodným způsobem; v případě díla zpřístupněného podle § 18 odst. 2 strojově čitelnými prostředky.

(3) Ustanoveními odstavců 1 a 2 není dotčeno ustanovení § 39d.

§ 39d: *Licence k rozmnožování díla pro účely automatizované analýzy textů nebo dat k vědeckému výzkumu*

Do práva autorského nezasahuje

a) 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, nebo

b) instituce kulturního dědictví, zhotoví-li pro účely vědeckého výzkumu rozmnoženinu díla za účelem automatizované analýzy textů nebo dat v digitální podobě, prováděné za účelem získání informací, zahrnujících mimo jiné vzory, tendence a souvztažnosti; takto zhotovenou rozmnoženinu je povinna uložit s vhodnou úrovní zabezpečení a může ji uchovávat pro účely vědeckého výzkumu, včetně ověření výsledků výzkumu.

53 Viz <https://www.zakonyprolidi.cz/cs/2000-121>.

54 Viz <https://eur-lex.europa.eu/legal-content/CS/TXT/?uri=CELEX:32019L0790>.

I když se bude právní praxe ve výkladu uvedených pasáží teprve ustalovat, zjednodušeně můžeme uvedené paragrafy chápat následovně: badatel (popř. badatelský tým) může pro svůj výzkum (automatizovanou analýzy textů nebo dat) pořídit kopii libovolných autorských děl, ale nesmí do nich zahrnout díla, u nichž si autor takové využití nepřeje.[55] Po dokončení výzkumu musí takto shromážděná díla nebo data odstranit. Instituce, které se zabývají výzkumem (tj. zejména vysoké školy), mohou také vytvářet kopie autorských děl pro účely automatizované analýzy textů nebo dat, ale na výsledky takového výzkumu se kladou některá omezení: zejména mají mít všechny další subjekty k takovým výsledkům stejný přístup a případný zisk se investuje do dalšího výzkumu. Výjimečné postavení mají instituce kulturního dědictví, tj. archivy, veřejné knihovny a muzea.[56] Pokud budou kopie autorských děl pro účely výzkumu dostatečně zabezpečené, mohou je pro tyto účely dlouhodobě uchovávat, mj. i pro ověření výsledků předchozího výzkumu.

6. DIGITÁLNÍ KNIHOVNY

České digitální knihovny zpřístupňují digitalizované publikace jednotlivých institucí prostřednictvím opensourcové aplikace Kramerius.[57] I když převažují monografie a periodika, využívá se také pro zpřístupnění starých tisků, map, hudebnin, grafik, zvukových a dalších typů dokumentů včetně tzv. borndigital[58] 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). U některých dokumentů jsou dostupná i metadata popisující logické části dokumentů, jako jsou kapitoly nebo články.

Z hlediska objemu dat a jejich struktury (vyváženosti) mají digitální knihovny blízko k webovým korpusům. Na rozdíl od webových stránek obsahují díla starší (de facto od počátků knihtisku) a autorsky chráněná, na druhou stranu v nich najdeme málo obsahu vytvářeného tak širokým spektrem tvůrců jako na internetu.

Všechny základní knihovní jednotky (monografie, periodika ap.) a jejich strany jsou v systému Kramerius identifikovány pomocí jedinečného perzistentního identi-

55 Pro tyto účely vznikne nejspíš specializovaný seznam děl, podobně jako v případě děl nedostupných na trhu, viz <https://sdnnt.nkp.cz/sdnnt/home>.

56 Viz např. záznam přednášky *Výjimky pro instituce kulturního dědictví dle směrnice o autorském právu na jednotném digitálním trhu*, která se uskutečnila krátce po schválení evropské směrnice; dostupné z <https://invenio.nusl.cz/record/407829?ln=cs>.

57 Zdrojové kódy: <https://github.com/ceskaexpedice/kramerius/>.

58 Jedná se o publikace, které vznikly v elektronické podobě a jejich tištěnou („papírovou“) verzi lze považovat za sekundární.

fikátoru UUID,[59] který je obvykle součástí URL adresy prohlíženého dokumentu, popř. metadat k němu.[60]

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.[61]

Přístup veřejnosti k obrazovým a textovým datům závisí na aktuálně platné autorskoprávní legislativě. V současné době pracuje systém Kramerius čtyř největších digitálních knihoven (Národní knihovna ČR, Moravská zemská knihovna v Brně, Knihovna Akademie věd ČR, Studijní a vědecká knihovna v Hradci Králové) se třemi způsoby zpřístupňování svého obsahu (Richter, 2020):[62]

— 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.
— 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, která vlastní fyzický exemplář daného díla. Volně dostupná jsou pouze popisná metadata a náhledy stran.

Díky nejnovějším změnám v AZ (viz výše) by takové rozlišování autorských děl pro potřeby automatizované analýzy dat nemělo hrát roli. Badatel by měl mít nárok na přístup ke všem digitálním kopiím, u nichž autor toto specifické využití výslovně nezakázal. Ale i podle staršího znění bylo možné v rámci digitálních knihoven poskytovat volná díla k libovolnému užití, kdy uživatel neužívá dílo dehonestujícím způsobem nebo si nenárokuje autorství díla. Za volná se považují autorská díla, u nichž

59 Univerzální jedinečný identifikátor (Universally Unique Identifier).

60 Např. publikace R. Kiplinga Můj sluha pes je dostupná z URL <https://www.digitalniknihovna.cz/mzk/uuid/uuid:a0e09900-2133-11e4-8413-5ef3fc9ae867>, kde poslední údaj v adrese („uuid:a0e09900-2133-11e4-8413-5ef3fc9ae867“) představuje jedinečný identifikátor této monografie.

61 Viz <https://github.com/ceskaexpedice/kramerius/wiki/>.

62 V roce 2023 se k nim přidá také Studijní a vědecká knihovna Plzeňského kraje. Zároveň systém Kramerius ve verzi 7 umožní zavedení dalších licencí, které mohou přesněji vymezit možnosti nakládání s digitalizáty získanými díky různým licenčním podmínkám a smlouvám s majiteli autorských práv.

uplynulo alespoň 70 let od smrti autora, resp. posledního ze spoluautorů. 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 (Lehečka, 2022, s. 16).[63]

Jedna instance Krameria zpřístupňuje digitální kopie obvykle z jedné instituce.[64] Rozcestník vybraných knihoven s tímto systémem nese název Digitální knihovna (bez data).[65] Seznam všech provozovaných instancí včetně podrobnější údajů o poskytovaných datech (počty publikací a naskenovaných stran) lze najít v Registru Krameriů (bez data). 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.[66] Česká digitální knihovna je také národním agregátorem pro portál Europeana (bez data). Nejjednodušší způsob, jak může badatel zjistit, zda existuje digitální kopie publikace, o kterou má zájem, představuje Registr digitalizace (2017). Jedná se o centrální místo, kde lze nalézt metadata o dokumentech, které byly v Česku digitalizovány nebo jejichž digitalizace se plánuje. U každého dokumentu jsou rovněž odkazy na digitální kopii v konkrétní instanci Krameria.

6.1 DATA PRO BADATELE

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

63 V českých digitálních knihovnách Kramerius se rozlišuje mezi tzv. chráněnými a volnými díly (z hlediska AZ), ale u volných děl se neuvádí žádná licence, takže badatelé mohou být v nejistotě, jakým přesně způsobem mohou s volnými díly nakládat. Kupříkladu Bavorská zemská knihovna (<https://www.digitale-sammlungen.de/en/>) zpřístupňuje digitalizáty např. jako dílo bez licence, které lze využít pouze nekomerčně (viz <http://rightsstatements.org/vocab/NoC-NC/1.0/>).

64 Výjimku tvoří případy, kdy paměťová instituce neprovozuje vlastní systém Kamerius, ale vyžívá instanci jiné instituce. A dále repozitáře Národní digitální knihovny a Moravské zemské knihovny, u nichž jsou na základě dohody některé publikace sdílené, takže se vyskytují v obou repozitářích současně; obě knihovny jsou totiž knihovny s celonárodním povinným výtiskem a mají do značné míry shodný fond. Obdobně některé knihovny získávají kopie digitalizovaných periodik z jiných knihoven, pokud je mají ve svém fondu. Je proto třeba vždy dávat pozor na to, z jaké instituce digitalizát reálně pochází.

65 Přístupné jsou zde pouze takové knihovny, které provozují systém Kramerius ve verzi 5 a vyšší. Aplikace Digitální knihovna zajišťuje jednotný přístup: každá knihovna má samostatnou webovou adresu, ale uživatelské rozhraní je pro všechny stejné.

66 Od Digitální knihovny se liší zejména tím, že umožnuje hledání v několik Krameriích z jednoho místa; aktuálně je možné prohledávat v šesti zapojených knihovnách. V budoucnu se plánuje sloučení České digitální knihovny a Digitální knihovny do jedné webové aplikace.

— obrazová data (obvykle uložená ve formátu JPEG 2000 a zpřístupňována ve formátu JPEG);
— textová data k jednotlivým stranám (jako prostý text nebo jako strukturovaný soubor XML ve formátu ALTO;[67] ALTO se u dokumentů objevuje postupně od přelomu let 2011 a 2012);
— popisná a další metadata, např. ve formátech MODS,[68] Dublin Core,[69] FOXML (Fedora Object XML)[70] a JSONLD dle specifikace IIIF Presentation API v3.[71]

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.

Z hlediska AZ popisná a další metadata nejsou považována za autorské dílo, takže je lze poskytovat bez omezení. V případě, že metadata tvoří databázi, tj. soubor obsahuje údaje k několika položkám, je k jejímu využití potřeba souhlas pořizovatele databáze, tj. knihovny nebo instituce, která tyto údaje vytvořila (Lehečka, 2022, s. 16). Pokud se licenční podmínky u tohoto druhu údajů neuvádí, je potřeba s nimi pracovat, jako by šlo o dílo (databázi) chráněné autorským zákonem.

6.2 PRÁCE SE SYSTÉMEM KRAMERIUS

Systém Kramerius byl od počátku vyvíjen tak, aby umožnil čtení publikací v digitální podobě. 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.

Architektura systému Kramerius využívá princip vícevrstvé aplikace, komunikace (výměna dat) mezi jednotlivými vrstvami probíhá pomocí programového rozhraní REST API, které je zdokumentované v repozitáři projektu na platformě GitHub.[72] 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.

Rozhraní REST API lze nicméně využít k programové práci s digitálními publikacemi, jak o tom svědčí např. aplikace krameriusimages[73] nebo XProcPipeline.[74] Zatímco první zmiňovaný skript slouží pouze ke stažení obrázků ke konkrétní pub-

67 Viz <https://github.com/altoxml>.
68 Viz <https://www.loc.gov/standards/mods/>.
69 Viz <https://www.dublincore.org>.
70 Viz <https://wiki.lyrasis.org/pages/viewpage.action?pageId=66585857>.
71 Viz <https://iiif.io/api/presentation/3.0/>.
72 Viz <https://github.com/ceskaexpedice/kramerius/wiki/>.
73 Viz <https://github.com/michal-josef-spacek/App-Kramerius-To-Images>.
74 Viz <https://github.com/daliboris/DL4DH/tree/main/XProcPipeline>.

likaci, druhá aplikace je komplexnější a slouží ke stažení většiny dostupných metadat (MODS, FOXML) a dat (JPEG, ALTO) a k obohacení textových dat pomocí lingvistických nástrojů výzkumné infrastruktury LINDAT/CLARIAH-CZ.[75] Tento program byl jedním z prototypů, který vznikl v rámci projektu „DL4DH — vývoj nástrojů pro efektivnější využití a vytěžování dat z digitálních knihoven k posílení výzkumu Digital Humanities".[76]

6.3 PRÁCE S NÁSTROJI PROJEKTU DL4DH

Grantový projekt DL4DH byl dokončen v roce 2022, takže nástroje, které během řešení vznikly, jsou k dispozici teprve krátce, přičemž s jejich dalším rozvojem se počítá i v následujících letech.

Součástí řešitelského týmu byli vedle zástupců knihoven i odborníci z oblasti digitálních humanitních a sociálních věd, kteří se k vytěžování obsahu digitálních knihoven snažili přistupovat z pohledu potřeb svých oborů. Díky společnému úsilí vznikla sada nástrojů, která slouží pro snazší vyhledání, přípravu a sdílení dat k dalšímu výzkumu. Řešitelé vycházeli z předpokladů, že samotný výzkum, analýza a interpretace získaných dat bývají velmi specifické a liší se nejen mezi obory, ale i v závislosti na položené výzkumné otázce. Vyvinuté nástroje se proto zaměřily na přípravu (vyhledání) dat, jejich obohacení o lingvistické informace (lemmatizace a morfologická analýza, identifikace pojmenovaných entit) a na jejich následné hromadné stažení. V dalších fázích výzkumu už bude záležet na badateli, jakou metodu pro analýzu a vytěžování zvolí. Díky jazykově obohaceným datům však může být cesta k relevantním výsledkům snazší (software se například může při analýze řídit jednotlivými lemmaty, nikoli konkrétními slovními tvary).

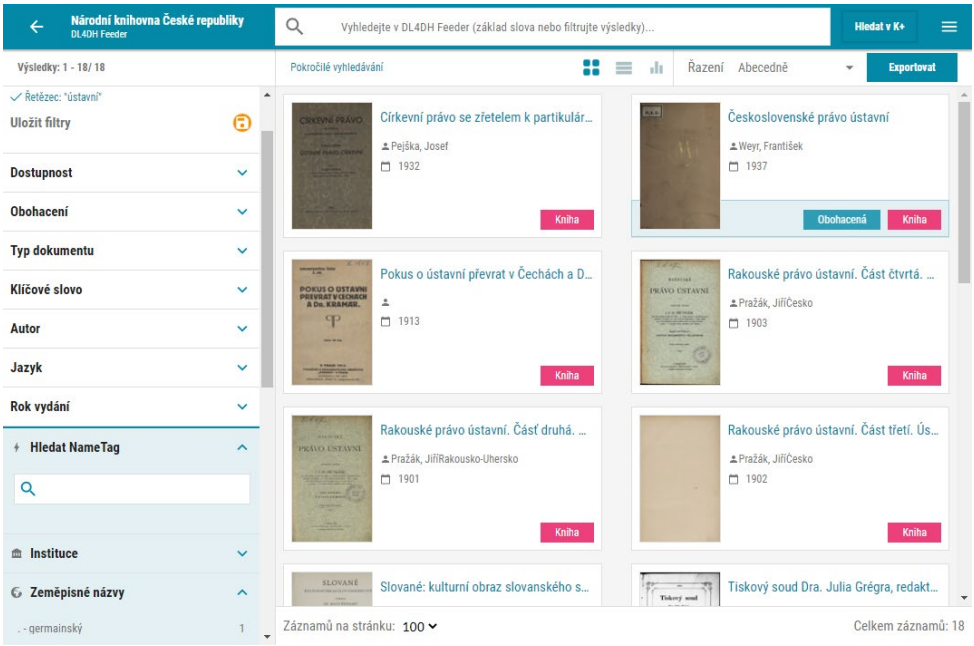
Pro obohacení o lingvistické informace se využívají nástroje vyvinuté a provozované v rámci infrastruktury LINDAT/CLARIAH-CZ: UDPipe 2 a NameTag 2 (viz výše). V této souvislosti je potřeba upozornit, že pokud bude badatel využívat ve svém výzkumu obohacující údaje dodané některým z těchto nástrojů, musí se řídit nejen licenčními podmínkami pro samotná autorská díla, ale i podmínkami, které do obohacených dat vnášejí výše zmiňované nástroje.[77]

Uživatelé mají díky nástrojům DL4DH přístup k datům v několika formátech. Textový obsah publikací je dostupný jako prostý text (zachycuje jednotlivé znaky kontinuálního textu bez formátovacích informací) nebo ALTO (zachycuje rozměry naskeno-

75 Aplikace je naprogramována v procedurálním jazyce XProc 3.0 (<https://xproc.org/specifications.html>), k jejímu spouštění slouží volně dostupná implementace procesoru XProc 3.0 v Javě MorganaXProc-III (<https://www.xml-project.com/morganaxproc-iii-se.html>) a bezplatná knihovna Saxon 10 HE (<https://github.com/Saxonica/Saxon-HE/tree/main/10>) pro transformace XSLT a XQuery.

76 Projekt NAKI II, č. DG20P02OVV002 <https://bit.ly/cmf-dl4dh-riv>, řešen v letech 2020–2022, řešitelské instituce: Knihovna akademie věd AV ČR, Moravská zemská knihovna v Brně, Národní knihovna ČR.

77 V případě uvedených aplikací se na obohacená data vztahuje licence Creative Commons BY-NC-SA, viz <https://creativecommons.org/licenses/by-nc-sa/4.0/>.



OBRÁZEK 1. Ukázka vzhledu DL4DH Feederu v testovací verzi

vané stránky, údaje o umístění textu na stránce a jeho formátování, a to jak na úrovni textových bloků, tak dílčích úseků) nebo JPEG (tj. obrázky v publikované kvalitě).

Metadata a obohacená textová data jsou k dispozici v několik strukturovaných formátech:

— CSV (comma-separated values)[78] a TSV (tab-separated values):[79] strukturovaná data uspořádaná do mřížky (v 1. řádku se nacházejí názvy sloupců);
— JSON:[80] zachycení dat, která mohou být organizována v polích nebo agregována v objektech;
— TEI:[81] standard konsorcia Text Encoding Initiative pro zachycení formálně i obsahově různorodých historických pramenů pomocí XML.

Hlavní komponentu, s níž bude pracovat badatel, který preferuje grafické uživatelské rozhraní, představuje webová aplikace DL4DH Feeder.[82] Její vizuální podoba vychází

78 Viz <https://datatracker.ietf.org/doc/html/rfc4180>.
79 Srov. <https://www.iana.org/assignments/media-types/text/tab-separated-values> a <https://www.loc.gov/preservation/digital/formats/fdd/fdd000533.shtml>.
80 Viz <https://www.json.org>.
81 Viz <https://tei-c.org/release/doc/tei-p5-doc/en/html/>.
82 Zdrojový kód: <https://github.com/LIBCAS/DL4DH-Feeder>, testovací instance je k dispozici zde: <https://feeder.nkp.cz>.

z webového klienta systému Kramerius[83] a je určena pro práci s daty uloženými v samostatném modulu Kramerius+[84] i digitální knihovně Kramerius. Mezi webovým klientem Krameria a DL4DH Feederem lze snadno přecházet pomocí tlačítka v záhlaví obou webů.

DL4DH Feeder umožňuje pracovat se všemi publikacemi, které jsou k dispozici v provázané instanci Krameria. Obohacené publikace jsou v uživatelském rozhraní označeny a uživatel s nimi může provádět některé specifické akce, např. vyhledávat v nich rozpoznané pojmenované entity, exportovat základní text včetně obohacujících údajů. Jelikož je proces obohacování náročný na výpočetní čas, předpokládá se, že vylepšené digitalizované publikace budou přibývat teprve postupně, v závislosti na požadavcích badatelů a personálních a technických možnostech digitalizačního pracoviště.

Pro plnohodnotnou práci s Feederem je nutné, aby se uživatel přihlásil, čímž získá přístup jednak k většímu množství funkcí, jednak k evidenci historie dotazů umožňující zjistit, zda v nastavení Feederu nebo v podkladových datech nedošlo od doby, kdy s aplikací 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. Při přihlášení musí uživatel souhlasit s podmínkami použití systému, v nichž jsou vymezena práva a povinnosti uživatelů.

Vyhledávat je možné v obvyklých bibliografických údajích (autor, titul), u obohacených děl funguje i pokročilejší prohledávání, např. v pojmenovaných entitách. Badatel může pomocí logických operátorů kombinovat různé podmínky. V uživatelském rozhraní se zobrazují fasety, které slouží k dalšímu zjemnění formulovaného dotazu na základě výběru hodnot k dostupným polím (nabídka hodnot se přizpůsobuje aktuálně vyfiltrovaným dokumentům).

Již ve webové aplikaci může badatel získat základní představu o zastoupení vyfiltrovaných publikací. Slouží k tomu grafy, u nichž lze na osu X umístit období vzniku dokumentu, jméno autora, klíčové slovo, typ dokumentu (kniha, noviny ap.) nebo jeho jazyk; frekvenční údaje na ose Y lze řadit vzestupně nebo sestupně.

Při exportu si badatel vybere požadované publikace: nejprve aktivací tlačítka „Exportovat" a následně označením relevantních dokumentů (buď po jednom, nebo všechny zobrazené položky na stránce naráz). Po této fázi následuje samotný export, kdy je potřeba zvolit formát (prostý text, ALTO, CSV/TSV, JSON, TEI) a u obohacených textů také údaje z obohacení, které se mají do výstupu uložit. Exporty provedené uživatelem jsou dostupné na samostatné stránce. Vzhledem k tomu, že se může jednat o velké objemy dat, budou se na serveru uchovávat pouze po omezenou dobu a záleží na badateli, aby data včas uložil na vlastní zařízení a použil je k dalšímu výzkumu.

Vedle webové aplikace s uživatelským rozhraním můžou badatelé zvolit cestu programového rozhraní a vyhledávat a exportovat data s využitím rozhraní REST

83 Zdrojový kód: <https://github.com/ceskaexpedice/kramerius-web-client>.

84 Aplikace tvořená dvěma databázemi a propojenými službami, která v projektu DL4DH zajišťuje mj. uložení a prohledávání obohacených dat k digitalizovaným publikacím.

API. Specifikace je dostupná na stránkách DL4DH Feederu,[85] vybrané příklady jsou k dispozici v repozitáři zdrojového kódu na GitHubu[86] a ukázka je dostupná také ve veřejném pracovním prostoru aplikace Postman.[87]

6.3.1 OMEZENÍ A PROBLÉMY

Jako většina aplikací, které pracují s historickými daty (v případě digitálních knihoven s publikacemi, které vznikaly v průběhu několika staletí, a zároveň s daty, tj. skeny a rozpoznáním, které vznikaly od roku 1992), je potřeba se i v případě nástrojů DL4DH vypořádat s touto proměnlivostí.

V prvé řadě je to nízká kvalita vstupních dat, tj. digitálních obrazů publikací, případně různá kvalita automatického rozpoznání textu.[88] To může být příčinou chyb při následných analýzách lingvistickými nástroji. Pokud u naskenovaného dokumentu není k dispozici rozpoznaný text ve formátu ALTO, nebude fungovat obohacení textu pomocí lingvistických nástrojů, které na tento typ vstupních dat spoléhají. V takovém případě je nutno kontaktovat kurátora sbírky a domluvit se na jejich rozpoznání pomocí aplikace pro OCR.

Na práci lingvistických nástrojů má dále vliv kombinace několika jazyků v textu. I když používané nástroje pro morfologickou analýzu (UDPipe 2), resp. pro rozpoznání pojmenovaných 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.

6.3.2 VÝHLEDY DO BUDOUCNA

Ostrá verze nástrojů DL4DH byla spuštěna teprve počátkem roku 2023, proto lze stejně jako u jiných projektů očekávat, že projde zatěžkávací zkouškou v podobě požadavků na přístup k velkým datům ze strany badatelů v oblasti humanitních a sociálních věd. Nahrávají tomu také změny v oblasti autorského práva, které umožňují využívat digitální kopie autorských děl pro automatizované analýzy, prováděné za účelem získání informací, zahrnujících mimo jiné vzory, tendence a souvztažnosti i v případech, že jsou díla autorsky chráněna (za splnění určitých podmínek).

Očekáváme jednak vylepšení funkcí zejména na straně uživatelského rozhraní, tj. DL4DH Feederu, jednak ve fázi přípravy dat ke zveřejnění. Může se na-

85 Viz <https://feeder.nkp.cz/swagger-ui/index.html>.

86 Viz <https://github.com/LIBCAS/DL4DH-Feeder>.

87 Viz <https://www.postman.com/winter-comet-20618/workspace/dl4dh-feeder>.

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

příklad jednat o zapojení aplikací pro OCR (např. PERO OCR (bez data)) pro rozpoznání textu v případě, že není k dispozici formát ALTO. Nebo využití novějších verzí nástrojů NameTag a UDPipe, jejichž aktualizace se v rámci infrastruktury LINDAT/CLARIAH-CZ připravuje.

7. ZÁVĚR

V oblasti humanitních a sociálních věd představuje významnou součást badatelské práce analýza primární a sekundární literatury. Vedle jazykových korpusů lze v posledních letech za vhodné zdroje písemných pramenů považovat digitální knihovny, které v českých zemích v letech 1992–2022 digitalizovaly přibližně 98,7 mil. stránek. Zpracování tak velkého objemu dat ze strany badatelů se stává možné díky rozvoji počítačových technologií, novým přístupům k datům i novým výpočetním metodám. Článek přibližuje dílčí zkušenosti ze zahraničí a přináší stručný přehled o zdrojích dat v českém prostředí. Zaměřuje se na nedávno dokončený projekt DL4DH, jehož cílem je nabídnout badatelům přístup k velkým objemům dat z digitálních knihoven Kramerius ve standardizovaných formátech (prostý text, ALTO, CSV/TSV, TEI, JSON) nejen prostřednictvím nové webové aplikace, ale i pomocí programového rozhraní REST API. Aby byla následná analýza publikací co nejsnazší, mohou být součástí stažených dat obohacující údaje z nástrojů UDPipe a NameTag, které vyvíjí a provozuje výzkumná infrastruktura LINDAT/CLARIAH-CZ.

LITERATURA

Lehečka, B. — Novák, D. — Kersch, F. et al. (2022): *Metodika přípravy dat z digitálních knihoven pro využití v digitálních humanitních vědách*. Knihovna AV ČR. Dostupné také z: http://invenio.nusl.cz/record/511549/files/Metodika_DL4DH.pdf

Richter, V. (2020): Zpřístupnění plných textů digitalizovaných knih a periodik prostřednictvím Národní digitální knihovny. *Informace — zpravodaj Knihovny AV ČR* [online]. (2) [cit. 2022-04-28]. 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/

Wilkens, M. — Ruan, G. (2020): *Geographic Locations in English-Language Literature, 1701–2011 (1.0). [Dataset]* [online]. 2020. HathiTrust Research Center [cit. 2023-01-19]. Dostupné z: https://doi.org/10.13012/2K5C-RF13

ELEKTRONICKÉ ZDROJE

Benko, V. (2014): Aranea: Yet Another Family of (Comparable) Web Corpora. In: P. Sojka — A. Horák — I. Kopeček — K. Pala (eds.), *Text, Speech and Dialogue* [online]. Cham: Springer International Publishing, Lecture Notes in Computer Science, s. 247–256 [cit. 2023-04-24]. Dostupné z: https://doi.org/10.1007/978-3-319-10816-2_31.

Benko, V. (2015): *Araneum Bohemicum Maximum: verze 15.04* [online]. Praha: Ústav Českého národního korpusu FF UK [cit. 2023-04-24]. Dostupné z: http://www.korpus.cz

Benko, V. (2020): *Araneum Bohemicum Maximum: verze 20.03* [online]. Bratislava: UNESCO Chair in Plurilingual and Multicultural Communication, Comenius University in Bratislava a [cit. 2023-04-24]. Dostupné z: http://unesco.uniba.sk

Bibliografie dějin Českých zemí [online] (2013). Praha: Historický ústav AV ČR [cit. 2023-04-24]. Dostupné z: https://biblio.hiu.cas.cz

Czech medieval sources FONTES [online] (2023). Praha [cit. 2023-04-24]. Dostupné z: https://sources.cms.flu.cas.cz

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

DALL·E 2 [online] (2022). San Francisco: OpenAI [cit. 2023-04-24]. Dostupné z: https://labs.openai.com

Digitální knihovna [online], bez data. Brno: Moravská zemská knihovna v Brně [cit. 2023-04-24]. Dostupné z: https://www.digitalniknihovna.cz

Europeana [online], bez data. [cit. 2023-04-24]. Dostupné z: https://www.europeana.eu/cs

HathiTrust: Digital Library [online] (2008–2023). [cit. 2023-04-24]. Dostupné z: https://www.hathitrust.org

HyperFontes: Metadatový modul databáze Czech Medieval Sources online [online] (2023). Praha: Centrum medievistických studií [cit. 2023-04-24]. Dostupné z: https://hyperfontes.cms.flu.cas.cz

Charles Translator for Ukraine [online] (2022). Praha [cit. 2023-04-24]. Dostupné z: https://lindat.cz/translation

ChatGPT [online] (2022). San Francisco: OpenAI [cit. 2023-04-24]. Dostupné z: https://chat.openai.com

Korpus českého verše [online] (2021). Praha: Ústav pro českou literaturu AV ČR, v. v. i. [cit. 2023-04-24]. Dostupné z: https://versologie.cz/v2/web_content/corpus.php

LINDAT/CLARIAH-CZ: Digitální výzkumná infrastruktura pro jazykové technologie, umění a humanitní vědy [online] (2023). Praha [cit. 2023-04-24]. Dostupné z: https://lindat.cz/cs

Manuscriptorium: Digital Library of Written Cultural Heritage [online] (2023). [cit. 2023-04-24]. Dostupné z: https://www.manuscriptorium.com/cs

Monasterium [online] (2023). ICARUS [cit. 2023-04-24]. Dostupné z: https://www.monasterium.net/mom/home

PERO OCR: demonstration application [online], bez data. Brno: Vysoké učení technické v Brně [cit. 2023-04-24]. Dostupné z: https://pero-ocr.fit.vutbr.cz

Registr digitalizace: Evidence dokumentů digitalizovaných v ČR [online] (2017). Praha: Národní knihovna ČR — Knihovna Akademie věd ČR — INCAD [cit. 2023-04-24]. Dostupné z: https://www.registrdigitalizace.cz

Registr Krameriů [online], bez data. Brno: Moravská zemská knihovna v Brně [cit. 2023-04-24]. Dostupné z: https://registr.digitalniknihovna.cz

Spoustová, J. — Spousta, M. (2012): *CWC2011*. Dostupné také z: http://hdl.handle.net/11858/00-097C-0000-0006-B847-6

Staročeská textová banka [online], bez data. Praha: Ústav pro jazyk český AV ČR, v. v. i., oddělení vývoje jazyka, verze dat 1.1.22 [cit. 2023-04-24]. Dostupné z: https://korpus.vokabular.ujc.cas.cz/first_form?corpname=STB-1.1.22.1

Straka M. — Straková, J. (2014): *NameTag* [online]. Praha: LINDAT/CLARIAH-CZ, digitální knihovna při Ústavu formální a aplikované lingvistiky, Matematicko-fyzikální fakulta Univerzity Karlovy [cit. 2023-04-24]. Dostupné z: http://hdl.handle.net/11858/00-097C-0000-0023-43CE-E

Straka M. — Straková, J. (2016): *UDPipe* [online]. Praha: LINDAT/CLARIAH-CZ, digitální knihovna při Ústavu formální a aplikované lingvistiky, Matematicko-fyzikální fakulta Univerzity Karlovy

[cit. 2023-04-24]. Dostupné z: http://hdl.handle.net/11234/1-1702

SYN [online], 2022. Praha: Ústav Českého národního korpusu FF UK [cit. 2023-04-24]. Dostupné z: https://www.korpus.cz/kontext/

SYN2020: reprezentativní korpus psané češtiny [online] (2020). Praha: Ústav Českého národního korpusu FF UK [cit. 2023-04-24]. Dostupné z: https://www.korpus.cz

Boris Lehečka | Moravská zemská knihovna v Brně |
Kounicova 65a, 601 87 Brno
ORCID: 0000-0003-4893-5537
lehecka@mzk.cz, boris@daliboris.cz