Marek Debnár, Marek Vician

Budovanie digitálnych textových zbierok 2.

(Teória a aplikácia)

Univerzita Konštantína filozofa v Nitre
2025

doc. Mgr, Marek Debnár, PhD.

Mgr. Marek Vician, PhD.

Vedeckí recenzenti:

doc. Mgr. Richard Změlík, PhD.

Prof. PhDr. Dalimír Hajko, DrSc.

Jazyková redakcia:

Gabriela Sedmáková

ISBN: 978-80-558-2290-7

Táto práca bola podporená Agentúrou na podporu výskumu a vývoja na základe Zmluvy č. APVV-20-0414.

AGENTÚRA NA PODPORU VÝSKUMU A VÝVOJA

Obsah

Korpusové a metodologické základy **4**
Digitálna zbierka ako výskumný objekt 4
Korpusy a ich typológia 6
Vyváženosť, reprezentatívnosť a selekcia 8
Literárny kánon v kontexte digitálneho výberu 11
Distant reading a kvantitatívne prístupy 13

Medzinárodné iniciatívy digitálneho výskumu literatúry **17**
Výskumná sieť COST Distant Reading for European Literary History 17
CLS INFRA a európska výskumná infraštruktúra 19

Slovensko a digital humanities **25**
Slovenský príspevok k ELTeC: korpus slovenských próz 25

Digitalizácia **29**
Zdrojový materiál: Selekcia, stav a rozsah 30
Skenovanie a spracovanie obrazu 31
Optické rozpoznávanie znakov (OCR) 33
Zásady OCR 33
Aplikácia OCR v projekte 33
Obmedzenia a výzvy OCR 34
Korekcia textu a zabezpečenie kvality 34
Manuálna a poloautomatická kontrola 35
Normalizácia 35
Kontrola verzií a dokumentácia 36
Opatrenia na zabezpečenie kvality 36
Výstupné formáty a príprava na kódovanie 36
Zhromažďovanie metadát 36
Typy zozbieraných metaúdajov 37
Metadáta podľa ELTeC usmernení 38

Kódovanie: štruktúrovanie textu podľa schémy ELTeC **39**
TEI a ELTeC 40
TEI 40
ELTeC 40
Štruktúra dokumentu TEI v ELTeC modifikácii 41
Vytvorenie TEI záhlavia 42
Opis súboru - `<fileDesc>` 43
Opis kódovania - `<encodingDesc>` 45
Opis profilu textu - `<profileDesc>` 45
Opis revízie - `<revisionDesc>` 46

Automatizácia tvorby metadát 46
Zdroje metadát 47
Metódy a nástroje automatizácie 49
Generovania TEI záhlaví v Dispre 50
Kódovanie tela textu 54
Anotovanie štruktúry v programe Microsoft Word 55
Transformácia anotovaného .docx súboru do XML 55

Prezentácia: Publikovanie pomocou TEI Publisher **59**
Základné princípy TEI Publisher-a 60
Oddelenie obsahu a prezentácie 60
Štandardizácia 60
Deklaratívna konfigurácia 60
Modelmi riadené transformácie 61
Okamžitá spätná väzba a itertívny vývoj 61
Užívateľská prívetivosť 61
Systémové nastavenie pre beh TEI Publisher-a 61
Výber hostiteľského prostredia 61
Inštalácia eXist-db 62
Nasadenie TEI Publisher-a 62
Adresárová štruktúra projektu 63
Konfigurácia projektu 63
Načítanie dokumentov 63
Náhľad a iterácia 63
Prispôsobenie rozhrania a správania publikovaného webu 65
Upravovanie rozvrhnutia a štýlov používateľského rozhrania 65
Konfigurácia navigácie a štruktúry dokumentu 67
Konfigurácia vyhľadávania 67
Rozšírenia dynamických schopností užívateľského rozhrania pomocou jazyka JavaScript 67
Prispôsobenie vykresľovania TEI elementov 68
Podpora viacjazyčného rozhrania a obsahu 68
Používateľské roly a spravovanie oprávnení 68

Bibliografia **71**

Korpusové a metodologické základy

Digitálna zbierka ako výskumný objekt

Už na úvod je potrebné rozlíšiť pojem digitálnej zbierky od pojmu korpusu, hoci sa často zamieňajú. Kým korpusy predstavujú systematicky zostavené súbory textov so zameraním na jazykovú analýzu a sú spravidla štruktúrované podľa lingvistických princípov (napr. obsahujú morfologickú alebo syntaktickú anotáciu), digitálne zbierky v literárnovednom výskume kladú dôraz predovšetkým na výber, reprezentáciu a interpretáciu textov ako kultúrnych objektov. Zatiaľ čo korpus má najčastejšie slúžiť ako nástroj na overovanie jazykových hypotéz, digitálna zbierka je formovaná s cieľom umožniť literárne, kultúrno-historické či estetické otázky. Inými slovami: zatiaľ čo korpus je predovšetkým nástrojom na extrakciu dát a je štruktúrovaný podľa lingvistických princípov (anotácie, frekvencie, morfologické značky), zbierka je aj metodologickým rámcom a epistemologickou výpoveďou o tom, čo považujeme za hodné výskumu. Digitálne zbierky v literárnovednom výskume kladú dôraz na výber, reprezentáciu a interpretáciu textov ako kultúrnych objektov. Nie sú tak len infraštruktúrou, ale aj výskumnou otázkou a metodologickým nástrojom.

Digitálna zbierka v kontexte digitálnych humanitných vied predstavuje zásadný posun v chápaní literatúry – nie ako uzavretého súboru výnimočných diel, ale systému, ktorý je dynamický a možno ho skúmať prostredníctvom výberu, usporiadania, spracovania a anotácie textov. Ide o posun od tradičného modelu individuálneho čítania a interpretácie k modelu, v ktorom je literatúra skúmaná ako štruktúra, kontext, distribučný priestor či historický vzorec. Zatiaľ čo v klasickom literárnom výskume zohrávali kľúčovú úlohu diela kánonizované na základe umeleckej hodnoty a kultúrnej prestíže, digitálne prístupy zdôrazňujú reprezentatívnosť, rozsah a transparentnosť výberu. V tejto súvislosti nadobúda pojem „zbierka“ nový význam – neoznačuje len súhrn textov, ale cieľavedome

vytvorený a spracovaný súbor, ktorý funguje ako výskumný nástroj. Digitálna zbierka nie je pasívny archív, ale aktívny analytický rámec: výber textov, spôsob ich digitalizácie, úroveň spracovania, typ anotácií a použité metadáta – to všetko ovplyvňuje, aké otázky možno klásť a aké odpovede môže výskum poskytnúť.

Zbierať literárne texty v digitálnej podobe preto neznamená ich len hromadiť alebo uchovávať. Na rozdiel od knižničného či archívneho chápania zbierky ako pasívnej databázy ide v prípade digitálnych zbierok o akt konštrukcie, ktorý si vyžaduje metodologické rozhodnutia a interpretačné zásahy. Digitálna zbierka nie je neutrálny výber, je zo kurátorský akt, ktorý zároveň odráža a formuje výskumné predpoklady.

V tomto prostredí prestáva byť literárny text výlučne umeleckým artefaktom a stáva sa zároveň dátovou jednotkou, ktorú možno analyzovať, porovnávať, zhlukovať či modelovať. Jeho hodnota nespočíva len v estetickej rovine, ale aj v jeho funkcii v rámci systému: ako reprezentant určitého štýlu, obdobia, témy alebo sociálno-historického kontextu. V dôsledku toho sa aj samotné zostavovanie digitálnej zbierky stáva epistemologickým gestom. To, čo je zahrnuté, čo je sprístupnené a čo je anotované, ovplyvňuje nielen predmet výskumu, ale aj jeho výsledky. Vznikajú tak zbierky, ktoré neodzrkadľujú literatúru „ako takú“, ale určitú jej modelovanú podobu – sprostredkovanú cez výberové kritériá, technologické možnosti a výskumné ciele.

Kvantitatívna zmena na vstupe – teda množstvo dostupných textov a objem dát – otvára širokú škálu výskumných otázok. Tieto sa týkajú jednak technických aspektov (ako prevod textov do digitálnej formy, typy digitalizačných nástrojov, automatické spracovanie), ale tiež konceptuálnych: ako sa rozhodovať pri zostavovaní tematických, historických či žánrových zbierok, ako zabezpečiť vyváženosť a čo presne má byť reprezentované. V tomto zmysle digitálna zbierka nie je to isté ako korpus. Hoci sa tieto pojmy prekrývajú, digitálna literárna zbierka je pluralitnejšia, flexibilnejšia a často špecificky zameraná na literárnovedné otázky: vývoj štýlov, distribúciu žánrov, sociálny profil autorov, dynamiku kánonu či kolísanie tém v čase.

Z pohľadu dištančného čítania (distant reading) predstavuje digitálna zbierka základný predpoklad novej výskumnej perspektívy. Kým podrobné čítanie (close reading) umožňuje hlbšie preniknúť do významovej štruktúry konkrétneho textu, dištančné čítanie umožňuje skúmať veľké množstvá textov naraz – a vďaka tomu odhaľovať štatistické vzorce, zoskupenia, výnimky či vývojové trendy, ktoré by pri tradičných metódach ostali neviditeľné.

Digitálna zbierka umožňuje sledovať literatúru ako systém, v ktorom sa dajú analyzovať frekvencie, rozptýlenia, podobnosti, klastre alebo štrukturálne odchýlky. Zostavenie takejto zbierky si však vyžaduje aj kritickú reflexiu samotného výberu. Otázka, ktoré texty reprezentujú dané obdobie, región alebo literárnu tradíciu, nemá jednoznačnú odpoveď. Tradične bol výber ovplyvnený hodnotiacimi kritériami (umelecká kvalita, inovatívnosť, kánonickosť). V digitálnom prostredí však môžu byť rozhodujúce iné faktory – frekvencia výskytu, prístupnosť, historická reprezentatívnosť a technická spracovateľnosť. Digitálne zbierky teda vytvárajú nové hierarchie a zároveň otvárajú priestor aj pre texty, ktoré by v tradičných výberoch nefigurovali.

Z tohto pohľadu sa vytváranie digitálnej zbierky ako výskumného nástroja stáva kľúčovým krokom pri budovaní výskumného rámca. Kvalita analýzy je totiž vždy závislá od kvality výberu – a výber nie je nikdy neutrálny. Zbierka ako výskumný objekt je tak zároveň objektom poznania aj prostriedkom poznávania. Formuje spôsob, akým sa literatúra chápe, analyzuje a prezentuje v digitálnom priestore. Digitálne zbierky zároveň prispievajú k demokratizácii výskumu. Umožňujú skúmať diela menej známych autorov a autoriek, regionálne či marginálne texty, rôzne jazykové podoby alebo žánre mimo hlavných prúdov. Ich prístupnosť a transparentnosť robí výskum opakovateľnejším, overiteľnejším a viac prepojeným s inými oblasťami humanitných štúdií a bádaní.

Napokon, digitálna zbierka nie je statický objekt ale otvorený rámec, ktorý možno aktualizovať, rozširovať a prehodnocovať. Ako taká sa stáva súčasťou infraštruktúry poznania, nie jeho doplnkom. V tom spočíva jej najväčší potenciál – nie v tom, že nahrádza tradičné čítanie, ale že rozširuje jeho možnosti a prehlbuje jeho základy.

Korpusy a ich typológia

Korpusy, ako systematicky vytvorené a digitálne spracované zbierky textov, predstavujú základnú infraštruktúru výskumu v oblasti digital humanities. Ich podoba, štruktúra a účel sa však výrazne líšia v závislosti od výskumných potrieb. Zatiaľ čo v jazykovede sú korpusy často zostavované s cieľom reprezentovať bežné jazykové použitie v rôznych komunikačných situáciách, literárna veda pracuje s korpusmi, ktoré sú zostavované podľa tematických, historických, žánrových alebo autorských kritérií.

Záujem o elektronické textové korpusy sa začal formovať už v 60. a 70. rokoch 20. storočia, predovšetkým v rámci jazykovedného výskumu v anglofónnom prostredí. Medzi prvé rozsiahle projekty patrí Brown University Standard Corpus of Present-Day American English (1964), ktorý slúžil ako základ pre vývoj prvých frekvenčných slovníkov. V 90. rokoch sa začali rozvíjať národné korpusy vo viacerých európskych krajinách a s rozšírením výpočtovej techniky sa vytvoril priestor pre špecializované a aj literárne korpusy. V súčasnosti ide o štandardnú infraštruktúru výskumu v oblasti lingvistiky, literárnej vedy a vývoja jazykových technológií.

Rozsiahle všeobecné textové korpusy, akým je aj hlavný korpus Slovenského národného korpusu (ďalej v texte SNK), sa zostavujú z textov rôznych štýlov a žánrov tak, aby čo najreprezentatívnejšie zastupovali jazyk, ako sa skutočne používa v písomných prejavoch. Ich cieľom je vytvoriť základný jazykový referenčný rámec, ktorý môže slúžiť ako zdroj empirických údajov pre rôzne výskumné účely – od štýlovej analýzy až po frekvenčné štatistiky. Na rozdiel od toho, v literárnovednom výskume sa často pracuje s menšími, cielene zostavenými korpusmi, ktoré vznikajú ako reakcia na konkrétnu výskumnú otázku. Takéto tematické alebo historické korpusy môžu byť zamerané na jedno obdobie, jeden žáner, skupinu autorov alebo konkrétny tematický rámec. Korpus ako digitálna zbierka nepredstavuje iba pasívny súbor textov – je výsledkom vedomej selekcie, kurátorskej činnosti a metodologickej reflexie.

Korpusy možno typologizovať podľa viacerých hľadísk. Z hľadiska pokrytia ide o referenčné korpusy, ktoré reprezentujú jazyk ako celok (napr. SNK), a výskumné korpusy, ktoré sú úzko profilované. Podľa zamerania ich možno rozdeliť na tematické, žánrové, autorské, historické, ale aj porovnávacie korpusy (napr. paralelné bilingválne korpusy). Z pohľadu spracovania rozlišujeme holé korpusy (plain text) a anotované korpusy, ktoré obsahujú informácie o morfológii, syntaxe, entitách, diskurzných značkách atď. Dôležitá je aj časová dimenzia: synchronické korpusy sa sústreďujú na jazyk v jednom časovom bode, kým diachronické umožňujú sledovať vývoj v čase. Ďalšou rovinou je vyváženosť: niektoré korpusy sú systematicky vážené podľa žánru, pohlavia autora alebo geografického pôvodu, iné sú skôr pragmatickým výberom dostupných textov.

Zostavenie korpusu je vždy výsledkom série rozhodnutí – nie iba technických, ale aj interpretačných. Každý z týchto rozhodovacích krokov má následky pre to, aké poznatky bude korpus generovať. Do akej miery má byť korpus reprezentatívny voči danej epoche?

Ktoré texty sú zahrnuté a ktoré vynechané? Sú všetky dostupné texty relevantné, alebo je potrebné vybrať len niektoré?

Príkladom historického literárneho korpusu je európsky projekt ELTeC (European Literary Text Collection), ktorý systematicky zhromažďuje beletristickú prózu z 19. a začiatku 20. storočia v rôznych európskych jazykoch. Tento typ korpusu je založený na konzistentných kritériách výberu: zastúpenie mužských a ženských autorov, časové pokrytie, národné reprezentácie a kontrola rozsahu. Cieľom je vytvoriť porovnateľné množiny textov, ktoré budú slúžiť nielen na kvantitatívnu analýzu, ale aj na vývoj a testovanie nástrojov NLP pre menej rozšírené jazyky.

Na národnej úrovni vznikajú aj špecializované projekty, akým je slovenský korpus DISPRO, ktorý dokumentuje prozaickú produkciu 20. storočia. V tomto prípade ide o autorsky orientovaný výber textov, ktoré sú zároveň spracované podľa štandardizovaných pravidiel a tvoria základ pre štýlometrické, tematické aj diachrónne analýzy.

Typológia korpusov teda nevzniká na základe formálnych rozdielov, ale na základe účelu a výskumného zámeru. Iný typ korpusu potrebuje ten, kto analyzuje vývoj syntaxe v publicistike, a iný ten, kto sleduje reprezentácie rodu v románe. Dôležité pritom je, aby výskumník rozumel vlastnostiam korpusu, ktorý používa - jeho obmedzeniam, výberovej logike aj spracovateľskému formátu. Práve táto transparentnosť a uvedomenie si limitov a možností robí z korpusu nielen nástroj, ale zároveň výskumný objekt.

Vyváženosť, reprezentatívnosť a selekcia

Otázky výberu textov, ich rozloženia v rámci literárneho korpusu a rozhodnutí, čo reprezentuje „literatúru ako celok“, patria medzi najkomplexnejšie výzvy digitálneho výskumu literatúry. Pri budovaní korpusu je rozhodujúce, čo bude zahrnuté, ako bude výber vyvážený, a prečo je ten-ktorý text považovaný za reprezentatívny. V tejto súvislosti zohráva zásadnú úlohu trojica pojmov: vyváženosť, reprezentatívnosť a selekcia.

Zostaviť literárny korpus znamená rozhodnúť sa, ktoré texty sú reprezentatívne a z akých dôvodov. Vytvoriť výber znamená nevyhnutne aj niečo vynechať. Reprezentatívnosť preto nie je daná objektívne – je to výskumné rozhodnutie, ktoré treba otvorene priznať a metodologicky zdôvodniť. Literárne diela sú v rámci výskumného výberu podrobené

redukcii – nie všetky texty možno zahrnúť, nie všetky sú dostupné, nie všetky majú rovnaký stupeň spracovania. To, čo sa považuje za reprezentatívne, závisí od cieľov výskumu a jeho interpretačného rámca. Z tohto pohľadu prestáva byť korpus len databázou – stáva sa nástrojom, ktorý modeluje literárne pole podľa určitej výskumnej logiky. Súbor textov je zostavený podľa výskumného zámeru, nie podľa existujúceho hodnotového rebríčka alebo historickej hierarchie diel. Literárne výbery v rámci DH nevyplývajú z náhodného výberu ani zo zaužívaných kánonických preferencií, ale sú dôsledkom výskumnej potreby a metodologických predpokladov. Otázka reprezentatívnosti sa tak nevzťahuje len na to, ktoré texty boli zahrnuté, ale aj na to, akým spôsobom boli spracované, akú úroveň anotácie dostali, a v akom formáte sú dostupné.

Korpusová vyváženosť nie je len štatistická kategória. Jej cieľom je zabezpečiť, aby boli zohľadnené rôzne vrstvy a podoby literárnej produkcie. Nie je účelom mať čo najviac textov, ale také, ktoré umožňujú zodpovedať výskumné otázky týkajúce sa literárnych noriem, žánrovej diverzity, autorskej identity alebo tematických preferencií. Zohľadňovanie výskumných potrieb znamená aj to, že reprezentatívnosť sa v jednotlivých výberoch líši. Čo je reprezentatívne pre výskum literárneho štýlu, nemusí byť reprezentatívne pre výskum tematickej zložky textu. Zbierka textov tak v konečnom dôsledku neodzrkadľuje literatúru „ako takú“, ale určitú predstavu o nej, sprostredkovanú cez metodologické rozhodnutia výskumníka. Preto sa aj reprezentatívnosť musí chápať ako konštrukt – ako niečo, čo je utvárané v procese výskumu, nie ako daný fakt.

To, že výber nie je náhodný, ale zameraný, má svoje dôsledky aj pri interpretácii výskumných výstupov. Každá vizualizácia, každé kvantitatívne zistenie stojí na výbere textov, ktorý ho umožnil. Preto je dôležité, aby sa výberové kritériá nestali neviditeľnými. Transparentnosť je základným predpokladom korektného digitálneho výskumu. Už samotná forma korpusu – jeho metadátová štruktúra, výber autorov, určenie časového rámca – ovplyvňuje, aké otázky si môžeme klásť. Zostavenie výskumného výberu nie je čisto technická záležitosť – ide o koncepčný akt, ktorý má vedecké aj ideologické dôsledky. Reprezentatívnosť sa tak stáva epistemologickou otázkou – otázkou, čo môžeme o literatúre tvrdiť na základe dát, ktoré máme k dispozícii. Dôležitým faktorom, ktorý vstupuje do výberu textov, je ich dostupnosť a digitalizovateľnosť. Mnohé texty nie sú elektronicky dostupné, nemajú overenú edíciu, sú fragmentárne alebo sú zaťažené autorskými právami. Tieto faktory ovplyvňujú, čo sa do korpusu vôbec môže dostať – a čo bude trvalo vylúčené.

Zároveň platí, že výskum založený na digitálnych dátach by nemal predstierať, že ide o kompletný obraz literatúry. Ide vždy o výber – a tento výber je podmienený historicky, technologicky či inštitucionálne. Výber teda nie je len metodologickým rozhodnutím, ale aj historickým a ideologickým aktom. To, čo považujeme za reprezentatívne, nie je dané raz a navždy – mení sa to spolu s výskumnými otázkami, s prístupmi k literatúre a s dostupnosťou technológií. Aj preto je dôležité korpusy chápať nie ako uzavreté súbory, ale ako otvorené rámce, ktoré možno rozširovať, revidovať a prepájať s inými výbermi. Len tak sa môže reprezentatívnosť približovať ideálu pluralitného zobrazenia literárnej reality.

Bolo by skresľujúce povedať, že kánon ako ho poznáme, sa v prípade kvantitatívnych výskumov vracia. Situácia, ktorá vyvolala potrebu redukovať či nájsť istú vyváženosť v stále narastajúcich digitálnych zbierkach, nebola vyvolaná len technologickým spracovaním, ale samotnou literárnou produkciou. Mark Algee-Hewitt a Mark McGurl nás vo svojej práci Medzi kánonom a korpusom[1] upozorňujú na tento nárast číslami: v roku 1975 bolo v anglickom jazyku publikovaných 7 948 kníh, v roku 2015 ich už bolo 278 985, čo znamená viac než tridsaťpäťnásobok. Ako vybrať z tisícok a tisícok titulov tie, ktoré stoja za pozornosť a sú nositeľmi estetických hodnôt? Aký kľúč použiť, ak chceme zostaviť reprezentatívnu vzorku toho, čo je literatúra, respektíve jazyk (daného smeru, obdobia, žánru literatúry)?

Pojem „reprezentatívnosť“ je v tejto súvislosti problematický ale zároveň esenciálny. Je nutné položiť si otázku, či je úmernosť, ktorá platí pri malých množinách textov (napr. charakteristické vlastnosti určitej národnej literatúry, žánru alebo obdobia), modelom pre veľké množiny (napr. svetová alebo európska literatúra)?“ Takto položená otázka otvára problematiku mierky a mierka je zasa spojená s výberom. Pri práci s veľkými korpusmi sa ukazuje, že klasická literárna veda sa opierala o veľmi úzky výber diel. Kritika tradičnej selekcie v literárnej vede je výrazná už u Morettiho, pre ktorého pochopiť literatúru ako celok, znamenalo vedieť odstúpiť od kánonizovaných textov na vzdialenosť, z ktorej sú viditeľné iné vzťahy než tie, ktoré nám odhaľuje tradičná literárna interpretácia.

Aj preto sa v rámci dištančného čítania objavuje potreba prehodnotiť kritériá výberu textov a otvoriť ich oveľa širšiemu spektru. Literárna hodnota tak už nebude výlučným kritériom selekcie; namiesto nej príde funkčnosť v rámci systému a schopnosť textu reprezentovať určitý typ, vzorec, štruktúru. Táto kvantitatívna zmena na vstupe otvára

[1] Algee-Hewitt and McGurl, *Between Canon and Corpus.*

širokú škálu otázok, počínajúc od technických otázok spracovania a prevodu textov do digitálnej formy, cez problémy zostavovania jednotlivých tematických, historických a žánrových korpusov, až po sofistikované postupy, akými sú počítačové analýzy rozprávania. Jedna z kľúčových otázok vo všeobecnej rovine, ktorá reflektuje rozdiel v prístupe k čítaniu, je preto otázka úmernosti: Je úmernosť, ktorá platí pri malých množinách textov (napr. charakteristické vlastnosti určitej národnej literatúry, žánru alebo obdobia), modelom pre veľké množiny (napr. svetová alebo európska literatúra)?

Zároveň však platí, že takto zostavený korpus je nástrojom interpretácie, ktorý umožňuje identifikovať pravidelnosti: Kvantitatívny zlom potom spočíva v odhaľovaní pravidelností, pretože čím viac pravidelností budeme schopní pomocou digitálnych nástrojov odhaliť vo všetkých sférach textu, tým viac sa budeme približovať k sfére dosiaľ nepoznaného. Základnou otázkou selekcie teda je, či a ako je možné reprezentovať literatúru ako celok. Digitálny výskum ponúka nový spôsob, ako túto otázku zodpovedať, nie cez výber „najlepších“, ale cez analýzu „najtypickejších“, „najčastejších“, „najvplyvnejších“ – alebo aj „najzabudnutejších“ diel.

Literárny kánon v kontexte digitálneho výberu

Digitálny výskum literatúry, najmä v jeho kvantitatívnej podobe, spochybnil zaužívané predpoklady o výbere textov, ktoré „reprezentujú“ literárne dejiny. Na rozdiel od tradičných prístupov sa neorientuje výlučne na tzv. vysokú literatúru, ale usiluje sa analyzovať celé literárne pole – vrátane periférnych, menej známych či populárnych diel. V tejto súvislosti sa nanovo otvára otázka literárneho kánonu a jeho metodologickej úlohy v digitálnej ére. Teoretická reflexia metodologického pohybu v analýzach literárnych textov smerom k väčším celkom, t.j. rozsiahlym digitálnym zbierkam, sa od začiatku opierala o kritiku kánonu a kánonicity. Dôvodom okrem samotnej selektívnosti bola aj výsledná interpretácia, ktorú Ansgar Nünning spája s definíciou literárneho kánonu samého: literárny kánon obvykle neoznačuje len korpus literárnych textov, ktoré skupina nositeľov, napr. celá kultúra či subkultúra považuje za hodnotné, autorizuje ich a snaží sa o ich zachovanie, ale aj korpus interpretácií, ktorý stanovuje, aké významy a hodnotové predstavy sa s kánonizovanými textami spájajú, teda stanovuje určitý výkladový kánon.[2]

[2]Porovnaj Nünning, *Lexikon Teorie Literatury a Kultury.*

Literárny kánon dlho zohrával dôležitú úlohu v literárnej histórii a estetike. Bol vôdzkou v neprehľadnom svete literatúry, pomáhal určovať hodnotu diela, a tým ovplyvňoval a formoval naše vnímanie literárneho poľa. Určoval, ktorá z kníh prežije alebo zostane zabudnutá. Bol zároveň nástrojom literárnej kritiky, histórie a teórie, ktorá ním posvätené diela a ich autorov nazvala klasikou, aby sa k nim dookola vracala a iniciovala ich nové čítania a interpretácie.

V digitálnych humanitných vedách sa preto začala formovať ostrá kritika voči tejto selekcii. Výrazným hlasom bol Franco Moretti, ktorý upozorňoval na metodologické obmedzenia spôsobené sústreďovaním sa na úzky kánon. Moretti si dobre uvedomoval aj spomínaný spoločenský a politický presah, avšak jeho výčitky voči kánonu sú skôr metodologické. Prvou je, že individuálne preferencie a konsenzus literárnych vedcov definujú celok na základe veľmi malého percenta textov označovaných ako kánonických. A druhá výčitka hovorí o tom, že texty vysokej umeleckej kvality, ktoré tvoria kánon, sú v rámci literárnej produkcie skôr výnimočné a zriedkavé, preto nemôžu korektne definovať literatúru ako celok. Odpoveďou na tento stav bol zber materiálu, t.j. rozširovanie digitálnych dát literárneho korpusu. Kritika kánonu však neznamená jeho úplné odmietnutie. V nových výskumných rámcoch má potenciál slúžiť ako jedna z referenčných vrstiev – nie však ako jediné meradlo hodnoty. Kvantitatívny výskum je z veľkej časti porovnávaním. Ak chceme odlíšiť to, čo je jedinečné, od toho, čo je bežné, potrebujeme vytvoriť reprezentatívnu vzorku, ktorej zostavenie je aktuálnou metodologickou otázkou. Práve v úvahách o tejto otázke sa objavuje miesto pre návrat kánonu. Nie je to však miesto na výslní, ako to bolo v minulosti, ale jedno z miest v rámci zložitejších štruktúr. Otázka, aký kánon zvoliť, alebo ako ho nahradiť, viedla k viacerým typom prístupov. Tie sa líšia podľa cieľov výskumu, povahy materiálu a metodologických východísk. Celkovo je možné identifikovať tri skupiny prístupov k tomu, ako zostaviť referenčnú vzorku. Prvou z nich je The "Best of…" alebo teda zoznam verejnosťou a kritikmi najoceňovanejších (kánonizovaných) kníh. Druhú skupinu reprezentujú ekonomické faktory. A posledný, tretí prístup sa zakladá na náhodnom výbere. V praxi digitálnych humanitných vied sa kánon čoraz častejšie určuje nie na základe estetiky, ale podľa overiteľných a kvantifikovateľných parametrov. Takýmto projektom je aj projekt Európskej zbierky literárnych textov ELTeC. Kánon je v tomto prípade reprezentovaný či určovaný počtom opakovaných vydaní, respektíve reedícií textu. Na Slovensku pod hlavičkou projektu Digitálna zbierka slovenskej prózy (ďalej v texte

používaná skratka DISPRO) vzniká lokalizovaná a rozšírená verzia, ktorá berie do úvahy špecifiká domáceho literárneho materiálu, a ktorej sa venuje aj táto publikácia.

Národné a jazykové špecifiká však do výberov vnášajú ďalšiu vrstvu komplexnosti. Najmä pri menších jazykoch a kultúrach platí, že výber nie je len technickým aktom, ale dotýka sa základných predstáv o tom, čo literatúra je – a čo ňou má byť reprezentované. A práve špecifiká domáceho literárneho materiálu sú tým, čo nie je možné vyčítať z kvantitatívnych dát. Napríklad totalitné režimy dvadsiateho storočia na našom území neumožňovali publikovať všetkým autorom rovnako, prípadne dané texty prechádzali cenzúrou a existujú tak vo viacerých verziách. Tieto znalosti sú však pri zostavovaní kvalitných literárnych korpusov kľúčové, ak teda chceme viac než len zlaté zrnká, ktoré nám občas prináša prúd slov. Vyvážená referenčná zbierka preto v podmienkach malých literatúr nemôže vychádzať len z počtu vydaní či medzinárodných štandardov, ako to predpokladá napríklad projekt ELTeC – je nutné ich lokalizovať. Kánon a kánonicita, ktorými sme sa zaoberali vo vzťahu k digitálnemu výskumu literatúry, výstižne ukazujú, ako digitálna súčasnosť nanovo definuje nielen pole skúmania literatúry, ale aj niektoré zaužívané literárno-estetické kategórie. Kánon má v tomto myslení mimoriadne postavenie, pretože bol od začiatku hlavným cieľom kritiky a v istom zmysle aj spúšťačom teórie dištančného čítania. V tomto kontexte prestáva byť kánon pevne daným zoznamom, ale stáva sa súčasťou širšieho procesu reprezentácie, ktorý zohľadňuje nielen početnosť výskytu či reedícií, ale aj kultúrnu rozmanitosť, tematickú pluralitu a historickú podmienenosť výberu. Najmä v prípade malých literatúr je digitálne zostavovanie výberov šancou zviditeľniť texty, ktoré by inak zostali na okraji, alebo by do kánonu nikdy neprenikli. Digitálny výber tak nemusí kánon nahrádzať, ale skôr ho revidovať a sprístupniť nové perspektívy na jeho vrstvy. V ére veľkých dát nie je otázka „kánon - áno alebo nie“, ale „ktorý kánon, ako vznikol a komu slúži“. Takéto uchopenie kánonu sa stáva nie pevným cieľom výskumu, ale jedným z nástrojov pri konštruovaní reprezentácie literárneho priestoru v dnešnom digitálnom veku.

Distant reading a kvantitatívne prístupy

Spomedzi inovácií, ktoré digitálne humanitné vedy vniesli do výskumu literatúry, patrí medzi najvplyvnejšie koncept distant reading (dištančného čítania), ktorý do literárnej teórie uviedol Franco Moretti. Tento prístup predstavuje zásadný obrat od tradičných kval-

itatívnych praktík - predovšetkým od tzv. close reading, teda podrobného, detailného čítania a interpretácie jednotlivých textov. Namiesto sústredenia sa na individuálne diela a ich významy, distant reading predpokladá prácu s rozsiahlymi súbormi textov, pričom cieľom nie je „čítať“ všetky texty, ale modelovať literárne systémy a odhaľovať makroštruktúry na základe kvantifikovateľných znakov a opakovateľných foriem. Spoločným menovateľom našich úvah bude už na začiatku spomínaný pojem Franca Morettiho dištančné čítanie (distant reading ďalej aj DR), ktorý reprezentujú makro- a mikroanalýzy veľkých elektronických textových korpusov. Nástroje, metódy a ciele, o ktoré v DR ide, sa síce opierajú o jazykovedné hľadisko, no nie sú primárne zamerané na výskum povahy jazyka, ale na odhaľovanie toho, ako a čo dané kategórie vypovedajú o ich vlastnej oblasti. Matthew L. Jockers v knihe Macroanalysis: Digital Methods and Literary History ilustruje vzťah medzi klasickým prístupom literárnych vedcov, tzv. podrobným čítaním (close reading) a DR ako rozdiel medzi ryžovaním a dolovaním zlata.[3] Toto obrazné prirovnanie vystihuje epistemologický posun, ku ktorému dochádza: výskum sa menej orientuje na výnimočné diela a viac na celkovú štruktúru literárneho priestoru, ktorú možno skúmať len pomocou agregovaných dát a ich analýzy. Moretti uvedený odklon rozšíril na celú množinu autorov a ich diel, ktorým literárna história venovala privilegovanú pozornosť: „Problém podrobného čítania vo všetkých svojich inkarnáciách (od nového kriticizmu po dekonštrukciu) sa zakladá na extrémne obmedzených kritériách. (…) Príliš veľa pozornosti venujeme jednotlivým textom, len preto, lebo si myslíme, že iba veľmi málo z nich je jej hodných.“[4] Týmto kritickým gestom Moretti tematizuje inštitucionálne slepé miesta literárnej vedy - predovšetkým jej redukcionizmus, pokiaľ ide o reprezentáciu literárneho poľa. DR reaguje na túto výzvu expanzívnym zberom dát a ich formalizovaným spracovaním. Moretti pri overovaní dát postupoval podobne ako Vladimír J. Propp. Aplikoval analytický prístup na literárnu vedu, avšak s využitím digitálnych technológií, čo mu umožnilo systematicky analyzovať omnoho väčšie množstvá textových dát. Podobne ako u Proppa, Viktora Šklovského a ďalších formalistov aj v Morettiho „kvantitatívnom formalizme“ je morfologická (žánrová) kategória predpokladom kvantifikačnej (makro)analýzy. Ďalším nevyhnutným aspektom dištančného čítania je už spomínaná vzdialenosť (distance), ktorá podmieňuje možnosť „čítať“ kvantitatívne dáta z hľadiska formy. [5] Historicky možno DR zaradiť do postformalizmu, v ktorom sa znovu aktualizuje štrukturalistické myslenie, no

[3] Jockers, *Macroanalysis.*
[4] Moretti, “Conjectures on World Literature” (s. 181)
[5] Moretti, *Graphs, Maps, Trees.* (s. 74)

rozšírené o nástroje veľkých dát a algoritmickej analýzy. Moretti vo svojich prácach, najmä v roku 2000, poukázal na limity tradičného podrobného čítania (close reading), ktoré sa zameriava na opakované čítanie úzkeho výberu literárnych diel, pričom ignoruje široké spektrum textov produkovaných v danom období, vrátane tzv. masovej literatúry. Podľa Morettiho literárni vedci neberú do úvahy „neviditeľné" časti literárneho poľa, čo môže viesť ku skreslenej predstave ako o literatúre, tak o jej vývoji.

Kritika exkluzívnosti kánonu je jadrom DR: Moretti usiluje o zviditeľnenie periférie, a to prostredníctvom kvantitatívne spracovaného, štatisticky reprezentatívneho výberu. Moretti vtedy svoju metódu opísal takto: „Už vieme ako čítať texty, teraz sa naučme ako ich nečítať. Dištančné čítanie, kde vzdialenosť je podmienkou poznania, nám dovoľuje zamerať sa na jednotky, ktoré sú oveľa menšie alebo oveľa väčšie ako text: literárne stvárnenie, témy, trópy – alebo žánre a systémy. A ak medzi veľmi malým a veľmi veľkým samotný text zmizne, je to jeden z tých prípadov, keď môžeme oprávnene povedať, že menej je viac."[6] Z tejto definície vyplýva, že ontologickou jednotkou skúmania v DR nie je text ako uzavreté dielo, ale dynamické formy v čase - konfigurácie žánrov, trendov, lexiky, ktoré tvoria sieťovú štruktúru literárneho systému. S výrazným nárastom textového materiálu – stonásobným až tisícnásobným v porovnaní s tradičnými súbormi – sa mení nielen rozsah výskumu, ale aj samotná výskumná prax. S takýmto objemom dát nemožno pracovať rovnakým spôsobom ako pri tradičných, limitovaných výberoch; kvantitatívna expanzia si vyžaduje nový prístup. Zmena mierky teda nie je len technologická, ale predovšetkým epistemologická: núti nás klásť si nové otázky a osvojiť si nové metódy, vrátane pokročilých foriem analýzy ako sú počítačové analýzy rozprávania. V tomto duchu pokračoval Franco Moretti aj vo svojom výskume svetovej literatúry. V článku „Style, Inc.: Reflections on Seven Thousand Titles" (British Novels, 1740–1850) z roku 2009 sa nezameriava na samotné texty, ale analyzuje názvy vyše 7000 románov – teda dáta, ktoré by tradičná literárna veda považovala za marginálne. Moretti si uvedomil, že „nateraz sú tituly najlepším spôsobom ako presiahnuť jedno percento románov, ktoré vytvárajú kánon, a uvidieť literárne pole ako celok" (Moretti, 2013, s. 181). Práve v tejto práci sa ukazuje potenciál kvantitatívnej štylistiky (štýlometrie) v literárnom výskume – nielen ako nástroja pre klasifikáciu štýlov, ale ako prostriedku na zachytenie štrukturálnych pravidelností v rozsiahlych korpusoch. Výsledky analýzy Morettiho tímu poukazujú na to, že aj paratextové údaje – ako sú tituly, metadáta, periodicita vydaní

[6] Moretti, *Distant Reading.* (s. 57)

či distribučné modely – môžu niesť významné sémantické a štrukturálne informácie. Ide o znaky, ktoré tradičná interpretačná prax často ignorovala, no ktoré v kontexte distant readingu nadobúdajú analytickú hodnotu ako nositelia systémových vzorcov. Svoj koncept literárneho poľa ako celku Moretti neodvodzuje z ničoho menšieho než z myšlienok o „svetovej literatúre." Morettiho vízia DR je globálna: nejde mu len o metodológiu pre národné literatúry, ale o rámec na mapovanie literárnych systémov ako historicko-kultúrnych totalít. DR vytvára kartografiu svetovej literatúry bez nutnosti jej úplného čítania.

Medzinárodné iniciatívy digitálneho výskumu literatúry

Výskumná sieť COST Distant Reading for European Literary History

COST akcia Distant Reading for European Literary History (CA16204), financovaná z programu Horizon 2020 Európskej únie, predstavovala prvý ambiciózny, celoeurópsky projekt zameraný na skúmanie literatúry z prespektívy digitálnych literárnych štúdií. Projekt prebiehal v rokoch 2017 až 2021 a reagoval na narastajúci záujem o výpočtové a štatistické metódy v humanitných vedách a potrebu pochopiť dejiny európskej literatúry nad rámec národných tradícií. Uplatňovaním princípov „vzdialeného čítania“ (distant reading) – prístupu, ktorý zaviedol Franco Moretti, a ktorý využíva výpočtovú analýzu na identifikáciu opakovaní a vzorcov v literárnych korpusoch – projekt ponúkol novú metodológiu na skúmanie kánonických aj nekánonických textov. Nešlo však iba o technický projekt, ale o intelektuálne a kolaboratívne úsilie, ktoré nanovo definovalo, ako možno literárne dejiny skúmať, interpretovať a vyučovať v 21. storočí.

Hlavným cieľom projektu bolo vyvinúť nástroje a zdroje pre celoeurópsku literárnu historiografiu, s dôrazom na viacjazyčnosť, inklúziu a reprodukovateľnosť. Projekt vychádzal z faktu, že tradičné národné kánony sú obmedzené a že vo väčšine európskych jazykov neexistujú porovnateľné literárne korpusy ako sú napríklad v angličtine. Cieľom projektu bolo preto vytvoriť podmienky pre komparatívny výskum založený na dátach, naprieč literárnymi kultúrami Európy. Od svojho začiatku bol projekt navrhnutý ako interdisciplinárna a medzinárodná spolupráca, do ktorej sa zapojilo viac ako 200 výskumníkov z vyše 30 krajín vrátane Slovenska. Široké disciplinárne zloženie účastníkov – od literárnych ved-

cov a korpusových lingvistov až po odborníkov na výpočtové spracovanie textu – umožnilo rozvíjať projekt súčasne na teoretickej aj metodologickej úrovni.

Najvýznamnejším výstupom projektu sa stalo vytvorenie European Literary Text Collection (ELTeC), viacjazyčného korpusu románov napísaných približne v období rokov 1840 až 1920. ELTeC predstavuje štruktúrovanú a vyváženú zbierku úplných textov románov vo viac ako tucte európskych jazykov, vrátane slovenčiny. Každý čiastkový korpus bol zostavený s cieľom zabezpečiť rozmanitosť z hľadiska pohlavia autora, dĺžky textu, žánru a kánonického statusu. Čiastočná spoluúčasť Slovenska na tvorbe tohto korpusu zabezpečila, že slovenská literatúra môže byť začlenená do komparatívneho a viacjazyčného výskumného rámca – čo umožňuje skúmať miestne texty v paralelnom vzťahu s väčšími literárnymi systémami, ako sú francúzsky, nemecký alebo anglický. Táto integrácia zároveň vedie k prehodnoteniu postavenia menších národných literatúr, ktoré už nie sú chápané ako periférne, ale sú vnímané ako plnohodnotná súčasť európskych literárnych dejín.

Všetky texty v ELTeC boli kódované v TEI-XML formáte podľa špecifickej schémy, vyvinutej v rámci projektu, ktorá zabezpečovala konzistentnosť medzi jazykmi, no zároveň umožňovala flexibilitu vo vzťahu k špecifikám jednotlivých literatúr. Cieľom bolo poskytnúť štrukturálne a jazykové informácie vhodné pre výpočtové spracovanie. Dôležité je, že celý korpus bol vytvorený v súlade s princípmi otvorenej vedy a všetky texty a metadáta boli sprístupnené prostredníctvom verejných repozitárov ako je GitHub. Projekt tým pádom naplnil zásady FAIR dát (Findable, Accessible, Interoperable, Reusable – nájditeľné, prístupné, vzájomne súčinné, znovu-použiteľné) a vytvoril infraštruktúru, ktorú môžu výskumníci a inštitúcie, vrátane tých v strednej a východnej Európe, ďalej využívať, upravovať a rozširovať.

Z metodologického hľadiska projekt vyvinul a rozšíril nástroje a pracovné postupy pre veľkorozmernú literárnu analýzu. Išlo napríklad o techniky štýlometrie, modelovania tém, priradenia autorstva, analýzy postáv a ich sietí, či sledovania sémantických a chronologických posunov. Tieto metódy boli sprostredkované prostredníctvom školení, workshopov a vedeckých stáží. Projekt tak významne prispel k rastu novej generácie výskumníkov v oblasti digitálnych literárnych štúdií.

Zásadným prínosom projektu však neboli len technické inovácie, ale aj teoretická reflexia. Projekt vytvoril priestor na diskusiu o kľúčových témach, ako sú formovanie literárneho

kánonu, rodová nevyváženosť v literárnej historiografii či jazyková a kultúrna politika európskej literatúry. Jedným z hlavných poznatkov bolo, že vzdialené čítanie nie je náhradou za čítanie zblízka, ale jeho doplnkom – a že umožňuje klásť nové výskumné otázky, ktoré by tradičné interpretačné prístupy neodhalili. Vzdialené čítanie nám napríklad umožňuje skúmať, ako sa žáner románu vyvíjal v slovenskej literatúre v porovnaní s inými tradíciami, alebo akým spôsobom boli (ne)zastúpené ženy-autorky v rôznych jazykových oblastiach. Projekt tak prispel k inkluzívnejšiemu a empiricky podloženejšiemu vnímaniu literárneho dedičstva. Aj po jeho formálnom ukončení v roku 2021 má projekt naďalej trvalý dosah. Jedným z hlavných následníckych projektov je CLS INFRA (Computational Literary Studies Infrastructure), opäť financovaný z programu Horizon 2020.

Z pohľadu výučby digitálnych humanitných vied predstavuje projekt Distant Reading modelový príklad toho, ako môžu rozsiahle, kolaboratívne a metodologicky hybridné iniciatívy transformovať výskum a aj pedagogiku vo všeobecnosti. Pre študentov a vedcov na Slovensku je dôkazom toho, aký význam má interdisciplinárna spolupráca, prepojenie technických a kritických kompetencií, ako aj aktívna účasť na formovaní otvoreného celoeurópskeho výskumu. Možnosť zahrnutia slovenskej literatúry do ELTeC korpusu navyše predstavuje významný krok smerom k dekánonizácii západoeurópskeho pohľadu a k potvrdeniu kultúrnej hodnoty stredoeurópskych literárnych tradícií v spoločnom európskom priestore.

CLS INFRA a európska výskumná infraštruktúra

CLS INFRA (Computational Literary Studies Infrastructure) je iniciatíva financovaná z programu Horizon Europe, ktorá bola spustená v marci 2021. Predstavuje nadnárodné konzorcium popredných európskych inštitúcií, medzi ktoré patria napríklad Poľská akadémia vied, Univerzita v Postupime, Rakúska akadémia vied, Karlova univerzita v Prahe, Univerzita v Gente, Trinity College v Dubline a ďalšie. Spoločným cieľom iniciatívy je vybudovať robustnú, vzájomne spolupracujúcu infraštruktúru pre digitálne štúdium literatúry (clsinfra.io). V čase, keď sa digitálne literárne archívy rýchlo rozrastajú, ich rôznorodosť a fragmentárnosť – keďže ich obsahy sú v rôznych formátoch, majú rôzne štandardy metadát a úrovne prístupnosti – stále predstavujú značnú prekážku pre výskum. CLS INFRA sa snaží tieto problémy prekonávať tým, že spája kvalitné dáta, pokročilé an-

alytické nástroje a vzdelávacie programy, a tým podporuje nové spôsoby bádania v oblasti viacjazyčného literárneho dedičstva Európy.

Základom práce CLS INFRA je dôraz na princípy FAIR ako aj princípov CARE (Collective Benefit, Authority to Control, Responsibility, Ethics), ktoré predstavujú doplnok k technickým štandardom FAIR a kladú dôraz na etické a spoločensky zodpovedné zaobchádzanie s dátami. Zdôrazňujú, že najmä pri práci s dátami týkajúcimi sa marginalizovaných, pôvodných alebo menšinových komunít treba zohľadniť ich právo na kontrolu, kolektívny prospech, zodpovednosť výskumníkov a kultúrne špecifiká. Týmto sa zabezpečuje, že literárne korpusy budú nielen dostupné, ale aj eticky a prakticky pripravené na akademické využitie. Budovanie infraštruktúry sa realizuje modulárnym prístupom: teda skrz metódy zdieľania dát, katalógy nástrojov, harmonizáciu metadát korpusov, NLP (Natural Language Processing) a analytické nástroje, ako aj workshopy a tréningové školy. Medzi konkrétne výstupy patria katalóg literárnych korpusov a nástrojov CLSCor, nástroje na transformáciu rôznych dátových formátov, technické postupy postavené na XML-TEI, Python/R API, Docker kontajneroch a Jupyter Notebookoch.

V oblasti dizajnu korpusov a literárnych metadát CLS INFRA poskytuje štandardy a odporúčania, ktoré určujú, čo spĺňa kritériá zahrnutia do výskumu (žánre, jazyky), ako aj rozsah (časové obdobia, témy), úplnosť metadát a reprezentatívnosť. Ukážkovým príkladom je zbierka ELTeC (European Literary Text Collection), ktorá je zámerne zostavená a vyvážená podľa jazyka, pohlavia autorstva, dátumu publikácie, dĺžky textu a ďalších kritérií – nie je iba náhodným výberom dostupných diel.

Súčasťou infraštruktúry CLS INFRA je rozsiahla sada výpočtových nástrojov, ktoré výskumníkom umožňujú analyzovať literárne texty pomocou metód známych ako spracovanie prirodzeného jazyka (Natural Language Processing - NLP). NLP je interdisciplinárna oblasť medzi informatikou, lingvistikou a umelou inteligenciou, ktorá sa zaoberá tým, ako počítače „chápu“ a „spracúvajú“ ľudský jazyk.

Work Package 8 (Pracovný balík 8) v rámci CLS INFRA je zameraný práve na vývoj a implementáciu NLP metód špecificky prispôsobených pre literárne texty. Medzi hlavné techniky, ktoré sa tu používajú, patria:

- **Named Entity Recognition (NER)** – rozpoznávanie pomenovaných entít: Ide o automatizované vyhľadávanie a klasifikáciu konkrétnych pomenovaných prvkov v

texte, ako sú mená osôb, geografické názvy, dátumy, organizácie a pod. Napríklad v románe môže NER identifikovať a kategorizovať „Franz Kafka“ ako osobu, „Praha“ ako miesto a „1912“ ako dátum. V literárnej analýze toto pomáha sledovať napríklad geografické súradnice deja alebo sieť postáv.

- **Relational Extraction (REX)** – extrakcia vzťahov: Táto technika ide o krok ďalej ako NER. Nejde len o identifikáciu entít, ale o automatické zisťovanie vzťahov medzi nimi. Napríklad môže z vety „Gregor Samsa žil so svojou rodinou v byte“ vyvodiť, že medzi entitami „Gregor Samsa“ a „rodina“ existuje nejaký vzťah spolubývania alebo rodinného puta. V literárnom výskume to umožňuje modelovať zložité siete vzťahov.
- **Sentiment Analysis (SA)** – analýza sentimentu: Tento nástroj sa pokúša určiť emocionálny tón textu, teda či je pozitívny, negatívny alebo neutrálny. V bežnej praxi sa používa napríklad na hodnotenie recenzií produktov, ale v literárnej analýze môže slúžiť na skúmanie nálad jednotlivých kapitol, zmien v emocionálnom vývoji postáv alebo vnímanej atmosféry textu.
- **Aspect-Based Sentiment Analysis (ABSA)** – analýza sentimentu podľa aspektov: ABSA je pokročilejšia verzia analýzy sentimentu. Nehodnotí iba celkový tón textu, ale zameriava sa na konkrétne aspekty alebo témy a určuje, aký sentiment sa k nim viaže. Napríklad: ak postava v románe opisuje mesto ako „krásne, ale preľudnené“, ABSA môže zaznamenať pozitívny sentiment k estetike mesta a negatívny k hustote jeho osídlenia. Táto veľmi jemná analýza umožňuje výskumníkom detailnejšie porozumieť postojom a hodnotovým rámcom postáv či rozprávača.

CLS INFRA navyše poskytuje praktické nástroje, ktoré študentom a výskumníkom umožňujú začať s vlastnou digitálnou analýzou textov. Ide o:

- **Jupyter Notebooks:** Interaktívne dokumenty, ktoré kombinujú kód, textové vysvetlenia, grafy a výstupy analýz. Sú ideálne na výučbu a dokumentáciu výskumných procesov v humanitných vedách.
- **Demonštračné kódy a skripty v populárnych NLP knižniciach**:
  - **spaCy:** Ide o knižnicu napísanú v programovacom jazyku Python, ktorá sa používa na spracovanie prirodzeného jazyka (NLP). Umožňuje vykonávať rôzne úlohy, ako je napríklad rozpoznávanie pomenovaných entít (mená osôb, miest, organizácií), prevod slov na ich základný tvar (lematizácia) či analýzy vetných

štruktúr (parsovanie). Takéto nástroje pomáhajú počítaču lepšie porozumieť tomu, čo text znamená.

- **Flair:** Nástroj pre analýzu sentimentu, klasifikáciu textu a sekvenčné označovanie.
- **LangChain:** Pokročilý nástroj pre prepojenie jazykových modelov (ako ChatGPT) s vonkajšími dátovými zdrojmi a nástrojmi.
- **Mistral-8x7b:** LLM, ktorého výkon umožňuje generovať či analyzovať zložité textové štruktúry.
- **Nervaluate:** Nástroj na hodnotenie presnosti výstupov NER modelov – umožňuje výskumníkom merať kvalitu rozpoznávania entít v literárnych textoch.

Tieto nástroje sú voľne dostupné cez platformy ako GitHub a poskytujú možnosť nielen pasívne čítať texty, ale aktívne s nimi experimentovať – skúmať štruktúru románu, meniť analytické parametre, porovnávať rôzne texty či vizualizovať sieť postáv. V prostredí digitálnych humanitných vied tak CLS INFRA prispieva k prechodu od tradičného výkladu literatúry ku kombinovanému prístupu, ktorý spája interpretáciu s výpočtovou presnosťou a replikovateľnosťou.

Okrem toho CLS INFRA organizuje sériu tréningových škôl, keďže jedným zo základných pilierov projektu je budovanie kapacít. Tieto školy, Praha (2022), Madrid (2023), Viedeň (2024), sa venujú témam ako je TEI kódovanie (štandardizované označovanie textov pomocou XML), tvorba a využívanie CQL dopytov (jazyk na vyhľadávanie v lingvistických korpusoch), nástroje ako UDPipe (na automatickú gramatickú analýzu textov), štýlometria (štatistické porovnávanie štýlu textov), programovanie s korpusmi (napr. v jazyku Python), práca s Linked Open Data (prepojené open data dostupné online), replikovateľnosť výskumu prostredníctvom nástroja Docker (na vytvorenie tzv. kontajnerov – teda balíkov, ktoré obsahujú všetky potrebné súčasti výskumného prostredia, vrátane verzií softvéru, knižníc, dát a skriptov) a sieťovej analýzy. Tréningové školy poskytujú výskumníkom – a to najmä z menej financovaných oblastí – technické zručnosti a kontakty potrebné na vytváranie a analýzu vlastných korpusov.

Významným nástrojom podpory je aj program Transnational Access (TNA) Fellowship, ktorý financuje výskumné pobyty na partnerských inštitúciách. Štipendisti tam získavajú praktické skúsenosti s komponentmi infraštruktúry, tvorbou korpusov, modelovaním

metadát, ladením NLP nástrojov a stratégiami publikovania. Výsledkom je ich lepšia pripravenosť zdieľať získané know-how so širšou komunitou.

CLS INFRA taktiež aktívne prispieva do širšieho ekosystému digitálnych humanitných vied, napríklad prostredníctvom prezentácií na konferenciách ako DH Benelux (2024), ako aj publikovaním výstupov ako sú rôzne metodologické prehľady, nástroje na zdieľanie dát, analýzy deficitov zručností a pilotné štúdie o trendoch v oblasti CLS . Spojením dát, nástrojov, výučby a komunitného zapojenia CLS INFRA zásadne mení spôsob, akým sa dnes realizuje literárny výskum v Európe – od fragmentárnych dátových archívov smerom k zjednotenej a dynamickej kultúre výpočtových literárnych štúdií. Jej výsledky a pokračujúce aktivity predstavujú výborný príklad pre každú národnú digitálno-humanitnú iniciatívu vrátane tých, ktoré sa začínajú rozvíjať aj na Slovensku.

Slovensko a digital humanities

Slovenský príspevok k ELTeC: korpus slovenských próz

Aj v slovenskom prostredí sa metodologické posuny v oblasti digitálneho výskumu literatúry stávajú čoraz viditeľnejšími. Projekt Digitálna zbierka slovenskej prózy vznikol ako odpoveď na potrebu vytvoriť tematicky a štylisticky rozvrstvený, odborne spracovaný korpus slovenskej literatúry, ktorý by bol dostupný pre širší okruh bádateľov. Tento korpus má umožniť nielen čítanie a vyhľadávanie, ale predovšetkým hlbšie analýzy založené na digitálnych metódach. Otvára sa tak možnosť novým pohľadom na slovenskú literatúru, či už ide o jej žánrovú dynamiku, rozloženie tematických preferencií, naratívne štruktúry alebo prehodnotenie literárnych dejín. Využívanie počítačových metód pri analýzach rozsiahlych zbierok literárnych textov, medzi ktoré patria kvantitatívna analýza textu, štýlometria, rozpoznávanie autorstva či rozpoznávanie emócií, sa v posledných desaťročiach stále viac dostáva do centra záujmu interdisciplinárneho humanitného výskumu. Pozornosť sa mu venuje nielen v kontexte digitálnych humanitných vied, ale vo všetkých odboroch, ktorých výskum sa primárne opiera o interpretáciu textov, tradične označovanú ako close reading (literárna veda, historiografia, filozofia, etika, náboženské štúdie atď.). Centrálnym impulzom pre vznik domáceho korpusu bola účasť Slovenska v medzinárodnom projekte Distant Reading for European Literary History 1840–1920 (COST CA16204), ktorý sa zameriava na zostavenie European Literary Text Collection (ELTeC) – viacjazyčnej zbierky literárnych textov. Projekt zastrešuje sieť výskumných tímov z viac ako tridsiatich piatich krajín a jeho cieľom je vybudovanie digitálnej infraštruktúry, ktorá umožní porovnateľný výskum naprieč národnými literárnymi tradíciami. Základom zbierky ELTeC je výber približne sto románov z obdobia 1840–1920 z každej zúčastnenej krajiny. Vstup Slovenska do projektu bol o to dôležitejší, že do tohto momentu sme nedisponovali dostatočne rozsiahlym, štruktúrou a štandardmi vybaveným digitálnym korpusom slovenskej liter-

atúry 19. storočia. K dispozícii boli síce rôzne zdroje – od oficiálnych archívov (napr. digitálny archív SNK), cez verejné projekty na dobrovoľníckej báze (Zlatý fond SME) až po neoficiálne súkromné weby – no žiaden z týchto zdrojov nespĺňal kritériá potrebné na zaradenie do ELTeC: overený edičný základ, štandardizáciu textov, štruktúru v TEI a spracovanie metadát. Snahou tvorcov slovenskej zbierky preto bolo nielen doplniť absenciu v európskom digitálnom priestore, ale aj podnietiť reflexiu o tom, čo reprezentuje slovenskú literatúru 19. storočia v jej tematickej a ideovej pestrosti. Príprava zbierky znamenala nielen technickú prácu (OCR, čistenie textov, ručné korektúry, kódovanie), ale aj edično-kritické rozhodovania: ktoré vydanie textu zvoliť, či bol text neskôr upravovaný, či existuje jeho variant, či je verifikovateľný jeho autorský status. Projekt ELTeC určil formálne kritériá pre zaradenie textov: dielo musí byť fikciou, napísaným v pôvodnom jazyku (nie prekladom), vydaným medzi rokmi 1840–1920, v rozsahu minimálne desaťtisíc slov, ideálne knižne publikovaným. Uprednostňované sú diela, ktoré sú voľne dostupné, neviazané autorskými právami. Do úvahy sa berú aj diela publikované v periodikách, pokiaľ spĺňajú ostatné podmienky. Zároveň sa kladie dôraz na diachrónnu vyváženosť – diela sú rozdelené do štyroch období (1840–1859, 1860–1879, 1880–1899, 1900–1920), čo umožňuje sledovať zmeny v čase. V slovenskej časti zbierky sa zohľadnilo pohlavie autora (zastúpenie ženských autoriek min. 10 %, max. 50 %), dĺžka románov (min. 20 % krátke do 50 000 slov, min. 20 % dlhé nad 100 000 slov) a úroveň kánonického statusu (min. 30 % vysoko kánonizovaných, min. 30 % nekánonických). Tento výber nebolo možné spraviť len na základe zoznamov "povinného čítania". V mnohých prípadoch bolo potrebné vyhľadávať diela zriedkavo reeditované, alebo znova publikované len okrajovo. Slovenská zbierka je teda produktom nielen technickej, ale aj metodologickej a konceptuálnej práce. Korpus je spracovaný v štandarde TEI, čo umožňuje štruktúrnu anotáciu (kapitoly, dialógy, poznámky), ako aj systematické metadátovanie (autor, pohlavie, žáner, rok vydania, kánonický status). Tým sa zaisťuje interoperabilita s inými zbierkami ELTeCu. Zber a spracovanie textov poukázali na špecifiká slovenského literárneho poľa. Na rozdiel od veľkých literatúr s bohatými digitalizačnými tradíciami je v prostredí malých literatúr potrebné oveľa väčšie metodické úsilie. Nedostatok štandardizovaných edícií, chýbajúce reedície, slabá digitalizačná infraštruktúra spôsobujú, že prístup k dielam je fragmentárny. Aj preto sa ukázalo, že niektoré súkromné iniciatívy (napr. ukrajinský príklad) sú neraz efektívnejšie ako oficiálne inštitúcie. ELTeC mal za cieľ nielen digitalizovať literatúry, ale ich aj zmysluplne reprezentovať. To si vyžaduje prehodnotenie tradičnej hierarchie a kánonov.

Vyváženosť a reprezentatívnosť zbierky nie je daná objektívne, ale je konštruktom, ktorý vzniká v procese výberu. Slovenský príspevok k ELTeC ukazuje, že aj malá literatúra môže byť súčasťou medzinárodných výskumných iniciatív, pokiaľ sa v nej skĺbi odbornosť, technologické know-how a metodologická vynaliezavosť. Korpus stoviek slovenských próz z 19. a začiatku 20. storočia je tak nielen bázou pre ďalší výskum, ale aj výzvou k redefinovaniu toho, čo znamená reprezentovať literatúru v digitálnom veku. A aký je teda aktuálny stav zbierky ELTeC? Od novembra 2020 sa zverejňujú na platforme Zenodo jednotlivé národné zbierky, ktoré spĺňajú kritériá a obsahujú minimálne päťdesiat textov.1 Pôvodne malo ísť o kolekciu prvých desiatich národných literatúr, existujú však dôvody domnievať sa, že ambiciózna úloha pripraviť viac ako desať reprezentatívnych čiastočných zbierok s polovičným počtom románov bude splnená do konca roka 2020 (projekt sa skončil v novembri 2021). Aktuálny stav slovenskej a ukrajinskej zbierky ELTeC je taký, že slovenské aj ukrajinské zastúpenie v projekte COST sa usiluje napriek vyššie opísaným problémom s digitálnymi zdrojmi spracovať a pripraviť svoje národné verzie zbierok ELTeC. V prípade Slovenska prípravu výrazne sťažuje absencia kvalitných digitálnych zdrojov, teda profesionálne zostavenej fulltextovej zbierky slovenskej literatúry, ktorá by sa dala v tomto prípade využiť. Na jej vytvorenie vo verzii značne rozšírenej a prispôsobenej slovenským špecifikám bol na jeseň 2020 v rámci výzvy Agentúry pre vedu a výskum podaný spoločný projekt troch pracovísk, dvoch ústavov SAV a jedného univerzitného pracoviska. Podpora projektov ako tento by výrazne napomohla nielen k skvalitneniu digitálneho výskumu slovenskej literatúry, ale aj k zapojeniu slovenských vedcov do európskych a širších medzinárodných, výskumných infraštruktúr v rámci dištančného čítania a digitálnych humanitných vied.

Digitalizácia

Digitalizácia býva typicky prvým krokom v projektoch digitálnych humanitných vied, ktoré vychádzajú z tlačených alebo rukopisných zdrojov. Je to proces transformácie fyzických objektov na ich digitálne reprezentácie, ktoré možno ďalej spracovávať, analyzovať, kódovať a zdieľať v digitálnom prostredí. V kontexte tejto prípadovej štúdie, digitalizačná fáza vytvára základ, na ktorom spočívajú všetky následné aktivity kódovania, analýzy a publikovania.

Dôležitosť tohto kroku nemožno preceňovať. Digitalizácia nielenže zachováva fyzickú integritu starnúcich a často krehkých diel tým, že znižuje potrebu manipulácie s nimi, ale otvára aj dvere vedeckým skúmaniam, ktoré v analógovej oblasti neboli možné. Vhodne spracované texty v digitálnej podobe možno analyzovať, vyhľadávať v nich, anotovať, porovnávať, vizualizovať a prepojiť s korpusmi alebo znalostnými systémami.

Digitalizácia však nie je len mechanický proces skenovania a extrakcie textu. Vyžaduje si dôkladné zváženie povahy zdrojového materiálu, zamýšľaného použitia digitálneho produktu a noriem, ktorým musia digitálne texty zodpovedať. Najmä staršie tlačené romány predstavujú špecifickú výzvu vzhľadom na ich fyzické vlastnosti, tlačové konvencie a typografické zložitosti, ktoré prináša historická ortografia, ligatúry alebo poškodenie pôvodných strán.

Táto kapitola má tri ciele. Po prvé, poskytnúť podrobný prehľad procesu digitalizácie od fyzického skenovania až po vytvorenie strojovo čitateľného textu. Po druhé, poukázať na praktické rozhodnutia, ktoré sa musia urobiť v každej fáze, vrátane výberu nástrojov, návrhu pracovného postupu a stratégií na opravu chýb. A po tretie, ukázať, že digitalizácia nie je len technickou úlohou, ale vedeckou praxou, ktorá si vyžaduje interpretačné povedomie, metodologickú transparentnosť a záväzok dlhodobého uchovávania a interoperability.

Na konci tejto kapitoly by mal čitateľ chápať, čo digitalizácia zahŕňa v projekte digitálnych humanitných vied a ako pripravuje pôdu pre následné kroky kódovania do špecifického analytického formátu a webovej prezentácie pomocou platforiem, ako je TEI Publisher. Práve tu sa začína cesta od tlačeného papiera k štruktúrovaným dátam, ktorá znamená transformáciu literárnych diel do digitálneho kultúrneho záznamu.

Zdrojový materiál: Selekcia, stav a rozsah

Pred začatím akejkoľvek digitalizácie je nevyhnutné byť dôkladne oboznámený so zdrojovým materiálom. Výber románov na zaradenie do digitálneho korpusu nie je neutrálnou úlohou - je formovaný vedeckými cieľmi, praktickými obmedzeniami a kurátorskými rozhodnutiami. Proces identifikácie vhodných textov sa začal bibliografickým prieskumom. Ten zahŕňal nahliadnutie do národných bibliografií, knižničných katalógov a archívnych súpisov s cieľom zostaviť zoznam potenciálnych kandidátov. Uprednostňovali sa diela, o ktorých sa vedelo, že majú historický, literárny alebo jazykový význam. Osobitná pozornosť sa venovala románom, ktoré nie sú bežne dostupné v digitálnej podobe inde, s cieľom prispieť novými materiálmi do širšej zbierky európskych literárnych textov (ELTeC).

Po výbere titulov sa hlavným kritériom stal fyzický stav a dostupnosť pôvodných tlačených vydaní. Staršie knihy sú často náchylné na poškodenia, majú krehký papier, vyblednutú tlač a známky predchádzajúceho používania (ako poznámky, podčiarkovanie alebo poškodenie). Tieto faktory môžu skomplikovať proces skenovania a znížiť presnosť optického rozpoznávania znakov (OCR). V niektorých prípadoch sa preskúmalo viacero výtlačkov, aby sa určilo vydanie v najlepšom stave, alebo aby sa porovnali variantné čítania. Ak to bolo možné, vyberali sa prvé alebo skoré vydania, aby sa zabezpečila historická presnosť a textová autenticita.

Rôznorodosť vydavateľov, typografických konvencií a kvality tlače z tohto obdobia priniesla ďalšie komplikácie. Písma môžu obsahovať dnes už nepoužívané ligatúry, neštandardné diakritické znamienka alebo svojrázne rozloženie. Pri príprave materiálu na OCR bolo potrebné zohľadniť marginálie, vydavateľské ornamenty a hlavičky strán, pretože tieto prvky by mohli vniesť šum alebo narušiť automatické rozpoznávanie.

Okrem toho bolo dôležité zdokumentovať všetky aspekty zdrojového materiálu vrátane metadát, akými sú informácie o autorstve, názve, roku vydania, vydavateľovi a mieste vydania. Tieto údaje sa zhromažďovali v štruktúrovanej forme v digitálnych súboroch vo formáte JSON, aby sa podporil proces kódovania aj budúca dohľadateľnosť digitálnych textov. Zachovanie transparentného prepojenia medzi fyzickým artefaktom a jeho digitálnou náhradou je kľúčové pre umožnenie budúceho overovania alebo redigitalizácie.

Celkovo možno konštatovať, že fáza spracovania pramenného materiálu nebola len prípravná, ale aj interpretačná a metodologická. Rozhodnutia, ktoré sa tu prijali, formovali reprezentačný rozsah korpusu a ovplyvnili každú ďalšiu fázu pracovného postupu. To podčiarkuje opakujúcu sa tému v práci v oblasti digitálnych humanitných vied: aj tie najtechnickejšie procesy sú zakotvené v otázkach významu, výberu a vedeckého rámcovania.

Skenovanie a spracovanie obrazu

Po identifikácii a príprave zdrojového materiálu vstupuje proces digitalizácie do svojej najtechnickejšej fázy: skenovania a spracovania obrazu. Tá zahŕňa transformáciu fyzickej tlačenej knihy na súbor vysokokvalitných digitálnych obrazov, ktoré budú slúžiť ako vstupná surovina na extrakciu textu a ďalšie spracovanie. Hoci sa skenovanie môže na prvý pohľad zdať jednoduché, dosiahnutie výsledkov vhodných pre vedecké digitálne edície si vyžaduje starostlivú pozornosť venovanú vybaveniu, rozlíšeniu, formátom súborov a postupom následného spracovania.

Pri tomto projekte sa použili ploché skenery a špecializované knižné skenery v závislosti od stavu a väzby fyzických zväzkov. Knihy s krehkými chrbtami alebo vzácnymi väzbami sa skenovali pomocou knižného skenera s V-kolískou, aby sa minimalizovalo fyzické namáhanie a skreslenie. Robustnejšie vydania sa skenovali pomocou plochých zariadení s vysokým rozlíšením. Bez ohľadu na hardvér sa všetky skeny vykonávali s minimálnym rozlíšením 300 dpi (bodov na palec), pričom sa uprednostňovalo rozlíšenie 400 - 600 dpi, ak to vyžadovala kvalita textu alebo zložitosť stránky.

Osvetlenie, zarovnanie a umiestnenie stránky pri skenovaní je vždy veľmi dôležité, pre zabezpečenie toho, aby naskenované obrázky boli jasné, rovnomerne osvetlené a bez tieňov

alebo skreslenia. Operátori sledovali kvalitu obrazu v reálnom čase, kontrolovali zaostrenie, kontrast a úplnosť stránky.

Surové skeny boli uložené v bezstratovom formáte TIFF. Tieto hlavné snímky boli archivované a slúžili ako bezpečná referenčná sada pre všetky ďalšie kroky. Na pracovné účely sa vytvorili menšie odvodené obrázky vo formátoch JPEG alebo PNG, ktoré sa použili v pracovných postupoch OCR a v nástrojoch na online prehliadanie.

Každej naskenovanej stránke bolo pridelené jedinečné meno súboru, prepojené s bibliografickými metadátami a uložené v štruktúrovanom adresári zodpovedajúcom zdrojovému dielu. Táto systematická organizácia bola nevyhnutná pre budúce dávkové spracovanie.

Po naskenovaní prešli obrázky niekoľkými automatizovanými a manuálnymi krokmi spracovania, aby sa pripravili na OCR. Bežné vylepšenia zahŕňali:

- **Odstránenie skosenia:** Oprava mierneho pootočenia, ku ktorému mohlo dôjsť počas skenovania.
- **Orezanie:** Odstránenie okrajov stránok, okrajov lôžka skenera alebo nepodstatných okrajových oblastí.
- **Binarizácia:** Konverzia farebných obrázkov alebo obrázkov v odtieňoch sivej na čiernobiele s cieľom zlepšiť kontrast OCR, v prípade potreby s použitím adaptívnych algoritmov prahovania.
- **Redukcia šumu:** Vyčistenie škvŕn na pozadí a odstránenie artefaktov, ako sú prach, záhyby alebo presvitanie atramentu z opačnej strany.

V prípade starších alebo poškodených kníh je občas potrebná manuálna retuš na odstránenie poznámok alebo na doplnenie čiastočne vyblednutých znakov, ktoré by sa inak pri OCR stratili.

Kontrola kvality bola trvalou súčasťou skenovania. Náhodne vybrané strany z každého zväzku sa vizuálne skontrolovali a náhľady OCR sa použili na náhodnú kontrolu čitateľnosti skenov. Strany s chybami - ako sú rozmazané, čiastočne zachytené alebo málo kontrastné - boli podľa možností naskenované znova. Tento opakovaný proces zabezpečil, že konečný súbor obrázkov bol nielen vizuálne verný originálu, ale aj optimalizovaný pre strojovú čitateľnosť.

Skenovanie a spracovanie obrazu slúži ako základná vrstva digitálneho korpusu. Presnosť, jasnosť a konzistentnosť tejto fázy priamo ovplyvňujú kvalitu textových údajov, ktoré možno extrahovať a zakódovať. Rovnako dôležité je, že tieto digitálne faksimile s vysokým rozlíšením fungujú ako most medzi fyzickou knihou a jej počítačovou reprezentáciou, pričom zachovávajú vizuálny kontext textu pre budúce referencie.

Optické rozpoznávanie znakov (OCR)

Po skenovaní a spracovaní obrazu zdrojového materiálu je ďalšou kľúčovou fázou digitalizačného procesu optické rozpoznávanie znakov (OCR). OCR sa vzťahuje na výpočtový proces, ktorým sa vizuálne tvary písmen a symbolov na naskenovaných stránkach konvertujú na strojovo čitateľný text. Tento krok je kľúčový pre transformáciu statických obrázkov na digitálne texty, v ktorých sa dá vyhľadávať, ktoré sa dajú upravovať a analyzovať a ktoré tvoria základ pre následné kódovanie a prezentáciu.

Zásady OCR

OCR softvér analyzuje bitmapové obrázky textu a snaží sa identifikovať vzory, ktoré zodpovedajú známym znakom alebo glyfom. Tento proces sa opiera o rozpoznávanie vzorov, analýzu rozloženia a čoraz viac o techniky strojového učenia. OCR systémy sa musia vysporiadať so širokou škálou typov veľkostí písma, rozloženia stránky a stavov skenovaných dokumentov.

Moderné nástroje OCR, ako napríklad Tesseract, ABBYY FineReader a OCRopus, využívajú pokročilé modely trénované na rôznych korpusoch dokumentov za účelom interpretáciu tlačeného textu s čoraz väčšou presnosťou. Úspešnosť však môže výrazne ovplyvniť kvalita vstupných obrazov, ako aj povaha zdrojového materiálu (napr. vek písma, poškodenie stránky, neštandardný pravopis).

Aplikácia OCR v projekte

Pre túto digitálnu zbierku sme použili predovšetkým softvér Tesseract,[1] pričom pracovný postup pozostával z niekoľkých fáz:

[1] "Tesseract User Manual".

1. **Príprava pred OCR:** obrázky sa vyčistili, binarizovali a segmentovali, aby sa oddelili textové oblasti. To pomohlo mechanizmu vyhnúť sa rozpoznaniu ilustrácií, čísel strán alebo poznámok na okrajoch ako hlavného textu.
2. **Výber jazyka a modelu:** Použil sa predtrénovaný model pre slovenský jazyk.
3. **Spustenie OCR:** Obrázky sa spracovávali dávkovo pomocou nástrojov príkazového riadka. Výstup bol vo forme obyčajného textu (plain text)
4. **Korekcia po OCR:** Výstup OCR sa kontroloval a opravoval poloautomaticky. Na opravu systematických chýb (napr. „rn“ nesprávne prečítané ako „m“ alebo ligatúry ako samostatné znaky) sa používali regulárne výrazy, pravopisné slovníky. Často však bola nevyhnutná aj manuálna oprava.

Obmedzenia a výzvy OCR

OCR je stále proces náchylný na chyby, najmä v prípade historických textov. Faktory, ako napr:

- Neštandardizovaný pravopis
- Písma starého typu (napr. Fraktur)
- Nepravidelné rozloženie alebo marginálie
- Vyblednuté alebo poškodené stránky

Tieto môžu viesť k nesprávnym rozpoznaniam znakov, ktoré sa bez prísnej kontroly môžu preniesť do neskorších fáz projektu. OCR Procesy si teda stále vyžadujú účasť človeka pri overovaní textového výstupu.

Korekcia textu a zabezpečenie kvality

Po získaní surového výstupu z OCR musí text prejsť dôkladnou korekciou a zabezpečením kvality, aby sa mohol spoľahlivo použiť na analýzu alebo ďalšie spracovanie. Tento krok je nevyhnutný, pretože chyby OCR môžu stále ohroziť presnosť, konzistenciu a použiteľnosť digitálneho korpusu, najmä ak ide o historické alebo neštandardné tlačené zdroje.

Text generovaný pomocou OCR je často nedokonalý. Chyby môžu byť rôzne, od jednoduchých zámen znakov (napr. nesprávne čítanie “1” ako “l” alebo “rn” ako “m”) až po zložitejšie problémy, ako sú nesprávne umiestnené riadky, nesprávne interpretovaná

diakritika alebo vynechané či zdvojené celé slová. Ak sa tieto nepresnosti nekontrolujú, môžu výrazne skresliť význam textu a viesť k nespoľahlivým výsledkom v nadväzujúcich úlohách.

Preto je systematický prístup k oprave textu nevyhnutnosťou. Cieľom je nielen opraviť chyby, ale aj zabezpečiť, aby výstupy dodržiavali konzistentný štandard, ktorý uľahčí neskoršie fázy anotovania, transformácie a publikovania.

Proces korekcie možno rozdeliť do niekoľkých odlišných fáz, z ktorých každá je prispôsobená špecifickým úrovniam automatizácie:

Manuálna a poloautomatická kontrola

Výstupy OCR sa môžu kontrolovať manuálne pomocou textových editorov, ale aj špecializovaných aplikácií akou je napríklad Transkribus[2]. Editori porovnávajú naskenované obrázky s rozpoznaným textom, pričom priamo opravujú chyby alebo označujú nejednoznačné oblasti pomocou vopred dohodnutých konvencií.

Poloautomatickú korekciu je možné vykonať pomocou:

- **Regulárnych výrazov:** na identifikáciu systematicky opakujúcich sa chýb OCR, napríklad bežných zámen písmen.
- **Slovníkov a programov na kontrolu pravopisu:** slúžia na označenie neštandardných slov, čím pomáhajú odhaliť nesprávne rozpoznané tokeny.
- **Nástroje špecifické pre jednotlivé jazyky:** napríklad morfologické analyzátory alebo rozpoznávače pomenovaných entít, ktoré slúžia na odhalenie nepravdepodobných konštrukcií alebo nezrovnalostí.

Normalizácia

V prípadoch, keď texty obsahujú archaickú alebo nejednotnú ortografiu, môže byť použitý určitý stupeň normalizácie. Ide o transformáciu variantných pravopisov do štandardnej podoby, a to buď ako paralelná vrstva (pomocou TEI elementov `<choice>` alebo `<orig>`/`<reg>`), alebo ako priame redakčné rozhodnutie.

[2] *Transkribus - Unlocking the Past with AI.*

Kontrola verzií a dokumentácia

Všetky opravy je vhodné zaznamenávať a sledovať pomocou systémov kontroly verzií (napr. Git). To umožnilo spolupracovníkom kontrolovať zmeny, udržiavať históriu úprav a v prípade potreby ich vrátiť. Tam, kde to je možné, je dobré formalizovať usmernenia pre opravy, aby sa zabezpečila konzistentnosť medzi prispievateľmi.

Opatrenia na zabezpečenie kvality

Po vykonaní opráv je vhodné vykonať celý rad kontrol zabezpečenia kvality:

- **Kontroly konzistentnosti:** zabezpečenie jednotného kódovania úvodzoviek, zlomov odsekov alebo odkazov na strany.
- **Overenie znakovej sady:** overenie, či sú všetky použité znaky správne zobrazené.
- **Predbežné overenie značiek:** použitie odľahčených TEI šablón pred začatím konečného kódovania na otestovanie správania sa textu v XML dokumente.

Niektoré projekty môžu používať aj tzv. dvojité kľúčovanie - porovnanie dvoch nezávislých prepisov s cieľom zistiť nezrovnalosti. Táto metóda je síce nákladná, ale poskytuje najvyššiu spoľahlivosť, najmä v prípade kľúčových textov.

Výstupné formáty a príprava na kódovanie

Po korekcii a kontrole kvality sú texty exportované do súborov s UTF-8 kódovaním , štruktúrovaných so základnými indikátormi riadkov alebo odsekov. Tieto súbory tvoria základný vstup pre kódovanie textov do XML formátov.

Zhromažďovanie metadát

Často prehliadaným, ale zásadným aspektom každého digitalizačného projektu je zhromažďovanie a štruktúrovanie metadát. V kontexte digitálnej zbierky metadáta popisujú nielen bibliografickú identitu každého diela, ale aj jeho históriu digitalizácie, pôvod zdroja a technické aspekty.

Samotný digitalizovaný text síce tvorí jadro zbierky, ale až vďaka metadátam sa z nej stáva robustný vedecký zdroj. Vďaka nim sú možné:

- **Katalogizácia:** názvy, autori, dátumy vydania etc. umožňujú konzistentné odkazovanie a indexovanie.
- **Vyhľadávanie a objavovanie:** filtrovanie a vyhľadávanie textov na základe polí metadát, ako je jazyk, pohlavie autora alebo obdobie vydania.
- **Pôvod a dohľadateľnosť:** zdokumentovanie toho, odkiaľ zdroj pochádza, ako bol digitalizovaný a kto ho digitalizoval, zabezpečuje transparentnosť a reprodukovateľnosť.
- **Interoperabilita:** integrácia s inými digitálnymi zbierkami, knižničnými systémami a výskumnými infraštruktúrami.

Typy zozbieraných metaúdajov

Metadáta pre projekt sa zbierali v niekoľkých kategóriách:

Bibliografické metadáta

Informácie o pôvodnom tlačenom zdroji zahŕňali:

- Názov diela
- Autor (vrátane normalizovanej formy a identifikátorov kontroly autority - VIAF)
- Rok a miesto vydania
- Vydavateľ
- Poznámky k vydaniu (napr. prvé vydanie, revidované vydanie)
- Počet strán a čísla zväzkov (ak ide o viac zväzkov)

Metadáta digitalizácie

- Technické a procedurálne detaily procesu digitalizácie:
- Model a rozlíšenie skenera
- Dátum digitalizácie
- Formáty súborov a použité metódy kompresie (napr. TIFF, JPEG)
- Zodpovedný prevádzkovateľ alebo inštitúcia

Digitalizácia

- Kroky následného spracovania (napr. odstránenie skreslenia, orezanie, verzia nástroja OCR)

Textové metadáta

Kvalitatívne a opisné aspekty textu ako pohlavie autorstva, rozsah alebo obdobie vzniku diela.

Administratívne metadáta

Informácie týkajúce sa práv a licencií:

- Stav autorských práv
- Licencia na digitalizáciu (napr. CC-BY, public domain)
- Inštitúcia alebo projekt spravujúci digitálny súbor
- Záznamy o verziách a aktualizáciách

Metadáta podľa ELTeC usmernení

Keďže cieľom projektu je prispieť do Európskej zbierky literárnych textov (ELTeC), aj metadáta sa riadili špecifikáciami definovanými v dokumentácii ELTeC. Tým sa zabezpečila kompatibilita s ostatnými zbierkami a podporila sa štatistická porovnateľnosť medzi rôznymi jazykovými korpusmi.

Medzi kľúčové ELTeC polia patria: `title`, `author`, `gender`, `year`, `language`, `filename`, `wordCount`

Tieto polia sa nakoniec integrovali do hlavičky každého TEI súboru, čím by sa opisné metadáta priamo prepojili s textovým obsahom.

Kódovanie: štruktúrovanie textu podľa schémy ELTeC

Po digitalizácii textov a ich príprave prostredníctvom procesov skenovania, OCR a manuálnej korekcie, je ďalším krokom pri vytváraní digitálnej zbierky zakódovanie textov do štandardného značkovacieho formátu. V kontexte tohto projektu sa kódovanie riadi usmerneniami stanovenými Európskou zbierkou literárnych textov (ELTeC), ktorá je postavená na prispôsobenej podmnožine normy Text Encoding Initiative (TEI).

Kódovanie literárnych textov je viac než len technická operácia - je to forma modelovania, pri ktorej sa štrukturálne, bibliografické a naratívne vlastnosti textu explicitne reprezentujú prostredníctvom značiek. Usmernenia ELTeC definovaním obmedzeného súboru prvkov a atribútov TEI dosahujú rovnováhu medzi vyjadrovacou silou a praktickou jednoduchosťou. To umožňuje kódovať širokú škálu literárnych textov z rôznych jazykov a časových období koherentným a porovnateľným spôsobom.

Proces kódovania sa začína pochopením a následným uplatnením princípov štandardov TEI a ELTeC. Zahŕňa zachytenie metadát (ako sú informácie o autorovi, história publikácie a jej jazyk) a štrukturálne značenie (ako sú kapitoly, odseky, nadpisy a citácie nachádzajúce sa v samotnom texte diela). Konečný produkt tohto procesu je XML súbor, ktorý je validný vzhľadom na ELTeC schému a vhodný na prezentáciu na webe a ďalšie spracovanie metódami dištančného čítania.

V tejto kapitole podrobne opisujeme ako prebieha kódovanie surových digitálnych textov do TEI formátu kompatibilného v súlade s ELTeC špecifikáciami. Začneme prehľadom modelov TEI a ELTeC, potom sa venujeme kľúčovým zložkám dokumentu TEI, pričom osobitnú pozornosť venujeme konštrukcii TEI záhlavia a označovaniu textových štruktúr. Zameriavame sa aj na procesy validácie a zabezpečenia kvality a skúmame škálovateľnosť kódovania väčších zbierok prostredníctvom automatizácie.

TEI a ELTeC

Skôr než sa začneme zaoberať praktickými aspektmi kódovania, je nevyhnutné pochopiť koncepčné základy dvoch štandardov: Iniciatívy pre kódovanie textu (TEI) a jej špecializovanej podmnožiny, Európskej zbierky literárnych textov (ELTeC).[1] Oba slúžia ako rámce na reprezentáciu textov v štruktúrovaných, strojovo čitateľných formátoch, ale líšia sa rozsahom, komplexnosťou a účelom.

TEI

TEI je štandard na reprezentáciu textov v digitálnej forme, extenzívne používaný v humanitných vedách. Jeho jadrom je modulárny a rozšíriteľný XML slovník určený na zachytenie rôznych štruktúr a vlastností textov - od prózy a poézie až po drámu, rukopisy, slovníky a ďalšie. TEI poskytuje prvky na označovanie všetkého od metadát dokumentu, štrukturálneho členenia a bibliografických odkazov až po jazykové a kritické poznámky.

Význam TEI spočíva v jeho flexibilite, čo však môže byť výzvou pre projekty, ktoré vyžadujú jednotnosť a interoperabilitu v rozsiahlych zbierkach. Keďže TEI umožňuje mnoho rôznych spôsobov reprezentácie podobných javov, kódovanie sa môže v jednotlivých textoch značne líšiť, pokiaľ nie je obmedzené spoločným prispôsobením alebo podmnožinou.

ELTeC

Tu vstupuje do hry ELTeC, schéma vyvinutá pod záštitou akcie COST Distant Reading for European Literary History, ktorá definuje zjednodušený a jednotný kódovací profil pre prozaické texty.

ELTeC je implementovaný ako prispôsobenie TEI pomocou systému ODD (One Document Does-it-all) a zameriava sa na obmedzený súbor TEI elementov, ktoré pokrývajú štrukturálne a bibliografické prvky potrebné pre literárne korpusy. Patria medzi ne:

- **Štruktúra textu:** odseky, nadpisy, kapitoly, epigrafy
- **Bibliografické metadáta:** názov, autor, dátum vydania

[1] Podrobnejšie sa autori týmto štandardom venujú v prvom diele učebnice - Budovanie digitálnych textových zbierok 1. (Technologické minimum) - na stranách 37 až 47.

- **Redakčné a jazykové prvky:** citáty, cudzie slová, dôraz
- **Minimálna sémantická anotácia:** pomenované entity (miesta, osoby), naratívne štruktúry (napr. zmeny rozprávača)

Zúžením rozsahu možností značkovania ELTeC zabezpečuje lepšiu medzijazykovú porovnateľnosť, jednoduchšiu validáciu a zaškoľovanie nových anotátorov. Uľahčuje tiež konzistentnejšie vykresľovanie a analýzu textov pomocou nástrojov, ako je TEI Publisher alebo platforiem na analýzu korpusov.

Štruktúra dokumentu TEI v ELTeC modifikácii

Ako sme uviedli vyššie, dokument validný vzhľadom na ELTeC schému je postavený na všeobecnej architektúre definovanej v usmerneniach TEI, ale obmedzuje a zjednodušuje ju tak, aby vyhovovala potrebám štandardizovaných literárnych korpusov. Každý TEI dokument je pritom XML dokument s koreňovým elementom `<TEI>`, ktorý obsahuje záhlavie `<teiHeader>` a telo textu `<text>`.

`<teiHeader>` obsahuje metadáta o zakódovanom texte. V ELTeC modifikácii ide o bibliografické informácie, ako napríklad názov diela, autor, údaje o publikácii a informácie o procese kódovania. Pozostáva z týchto hlavných zložiek:

`<fileDesc>`: Poskytuje bibliografický opis textu vrátane názvu (`<titleStmt>`), informácií o publikácii (`<publicationStmt>`) a údajov o zdroji (`<sourceDesc>`). `<encodingDesc>`: Opisuje použité princípy kódovania, často s odkazom na usmernenia ELTeC alebo akékoľvek špecifické úpravy projektu. `<profileDesc>`: Obsahuje ďalšie popisné metadáta o obsahu, ako napríklad použité jazyky, počet slov alebo perspektíva rozprávania. `<revisionDesc>`: Zaznamenáva zmeny vykonané v dokumente a ponúka informácie o jeho verziách a pôvode.

Zatiaľ čo TEI umožňuje mimoriadne podrobné metadáta, ELTeC štandardizuje a zjednodušuje očakávaný obsah záhlavia, pričom uprednostňuje elementy, ktoré podporujú interoperabilitu a kvantitatívnu literárnu analýzu.

Element `<text>` obsahuje vlastný obsah literárneho diela a zvyčajne sa delí na tieto časti:

- `<front>` (nepovinné): Pre titulné strany, venovania, predslovy alebo úvody.

- `<body>`: Hlavný obsah textu zakódovaný v štruktúrovaných prvkoch
  , ktoré predstavujú rôzne typy členenia príbehu.
- `<back>` (nepovinné): Pre záverečné časti, ako sú dodatky, epilógy alebo slovníky.

V rámci elementu `<body>` je obsah štruktúrovaný pomocou prvkov, ako sú:

- `<p>`: Odseky prózy.
- `<head>`: Nadpisy sekcií alebo kapitol.
- `<q>`: Riadkové alebo blokové citácie.
- `<foreign>:` Cudzojazyčný materiál.
- `<hi>`: Zvýraznenie textu, napríklad prostredníctvom kurzívy.
- `<name>`: Označené pomenované entity, niekedy anotované atribútmi označujúcimi typ (napr. osoba, miesto atď).

Tieto prostriedky umožňujú reprezentáciu hierarchie pôvodného románu a zároveň ďalšie počítačové spracovanie, napríklad vyhľadávanie pomenovaných entít alebo mapovanie štruktúr kapitol.

Hlavným cieľom ELTeC špecifikácie nie je zakódovať všetky možné vlastnosti literárneho textu, ale nájsť rovnováhu medzi expresivitou a konzistentnosťou. Obmedzenie slovníka elementov a predpísanie jasných postupov kódovania uľahčuje manuálne a poloautomatické kódovanie textov, validáciu dokumentov a vykonávanie medzitextových porovnaní.

V nasledujúcich častiach sa bližšie pozrieme na každú z týchto zložiek v praxi, pričom začneme záhlavím dokumentu a potom sa budeme venovať náležitostiam tela textu kódovaných diel.

Vytvorenie TEI záhlavia

Element `<teiHeader>` je základnou zložkou každého dokumentu zodpovedajúceho schéme TEI a slúži ako jeho popisná a administratívna predná časť. V rámci ELTeC zohráva záhlavie kľúčovú úlohu pri zabezpečovaní konzistentnosti v celom korpuse a pre ľudských používateľov aj výpočtové systémy poskytuje informácie pre pochopenie kontextu kódovaného textu. V tejto časti uvádzame, ako zostaviť TEI záhlavie TEI v súlade s usmerneniami ELTeC, pričom venujeme pozornosť jeho povinným aj nepovinným zložkám.

Záhlavie TEI obsahuje metadáta, teda dáta o dátach. Uvádza bibliografické údaje, informácie o procese digitalizácie a kódovania a kontextové poznatky potrebné na na ďalšie využitie pri aplikácií metód dištančného čítania. Dobre zostavené záhlavie umožňuje:

- Kontrolu citácií a bibliografie
- Sledovanie zdrojov a pôvodu kódovaného textu
- Interoperabilitu medzi systémami a korpusmi
- Analýzu a filtrovanie v rozsiahlych štúdiách.

Opis súboru - `<fileDesc>`

Ide o povinnú časť, ktorá obsahuje bibliografický opis textu, pričom obsahuje nasledujúce elementy:

- `<titleStmt>`: Obsahuje elementy `<title>` (názov románu), `<author>` (informácie o autorovi alebo autoroch) a `<respStmt>` (vyhlásenie o zodpovednosti) s informáciami o kodérovi(-och).
- `<publicationStmt>`: Zvyčajne generický v ELTeC, často sa v ňom uvádza, že korpus je k dispozícii pod licenciou Creative Commons a pomenovanie projektu.
- `<sourceDesc>`: Opisuje zdroj, z ktorého bol digitálny text odvodený. Zvyčajne obsahuje bibliografické údaje (autor, názov, vydavateľ, dátum, vydanie) pôvodného tlačeného vydania alebo skenu.

Pre ilustráciu tejto časti uvádzame relevantný fragment záhlavia dokumentu jedného z dokumentov tvoriacich našu zbierku:

```
<tei:fileDesc>
 <tei:titleStmt>
  <tei:title>
   Márnosť všetkého : ELTeC edícia
  </tei:title>
  <tei:author ref="viaf:14806974">
   Slančíkova-Timrava, Božena (1867 - 1951)
  </tei:author>
  <tei:respStmt>
```

```
    <tei:resp>
     editor
    </tei:resp>
    <tei:name>
     Meno Editora
    </tei:name>
   </tei:respStmt>
  </tei:titleStmt>
  <tei:extent>
   <tei:measure unit="words">
    10911
   </tei:measure>
  </tei:extent>
  <tei:publicationStmt>
   <tei:p/>
  </tei:publicationStmt>
  <tei:sourceDesc>
   <tei:bibl type="printSource">
    <tei:author>
     Slančíkova-Timrava, Božena
    </tei:author>
    <tei:title>
     Márnosť všetkého
    </tei:title>
    <tei:pubPlace>
     Turčiansky Sv. Martin
    </tei:pubPlace>
    <tei:publisher>
     Kníhtlačiarsky účastinársky spolok
    </tei:publisher>
    <tei:date>
     1908
    </tei:date>
```

```
    </tei:bibl>
   </tei:sourceDesc>
  </tei:fileDesc>
```

Opis kódovania - `<encodingDesc>`

V tejto časti sú uvedené informácie o postupoch kódovania a dodržiavaných normách, alebo aj informácie o účele, za ktorým bol dokument vytvorený:

```
<encodingDesc n="eltec-1">
  <projectDesc>
    <p>Kolekcia textov pre použitie v APVV projekte Digitálna zbierka slovenskej
  prózy, Marec 2025.</p>
  </projectDesc>
</encodingDesc>
```

Opis profilu textu - `<profileDesc>`

Táto časť obsahuje popisné metadáta o obsahu samotného románu a môže obsahovať tieto informácie:

- `<langUsage>`: opisuje jazyky, podjazyky, registre, nárečia atď. zastúpené v texte.
- `<textDesc>` : poskytuje opis textu z hľadiska jeho situačných parametrov.

Opäť na ilustráciu uvádzame fragment dokumentu nachádzajúci sa v našej zbierke:

```
  <tei:profileDesc xmlns:eltec="http://distantreading.net/eltec/ns">
   <tei:langUsage>
    <tei:language ident="slk"/>
   </tei:langUsage>
   <tei:textDesc>
    <eltec:authorGender key="M"/>
    <eltec:size key="short"/>
    <eltec:canonicity key="unspecified"/>
    <eltec:timeSlot key="T4"/>
```

Kódovanie: štruktúrovanie textu podľa schémy ELTeC

```
   </tei:textDesc>
  </tei:profileDesc>
```

Opis revízie - <revisionDesc>

Protokol zmien a aktualizáciách vykonaných v súbore. Tieto informácie podporujú transparentnosť a atribúciu zodpovednosti pri spoločných alebo dlhodobých projektoch.

```
<revisionDesc>
  <change when="2025-04-15" who="#mv">Úvodné kódovanie.</change>
</revisionDesc>
```

Automatizácia tvorby metadát

V kontexte kódovania rozsiahlych zbierok je ručné zadávanie metadát pre každý dokument nielen pracné, ale aj náchylné na chyby. Automatizácia tvorby TEI záhlavia - najmä opakujúcich sa alebo štruktúrovaných metadát - môže výrazne zvýšiť efektivitu tvorby korpusu. Ak pracujete s desiatkami alebo stovkami textov, opakované vytváranie tej istej štruktúry - vypĺňanie bibliografických polí, údajov o projekte a kódovanie informácií - môže zabrať nespočetné množstvo hodín. Automatizácia výrazne znižuje túto réžiu tým, že generuje validné TEI záhlavia z už existujúcich štruktúrovaných údajov, čím uvoľňuje ľudskú pozornosť pre úlohy, ktoré si vyžadujú intelektuálnu prácu, ako je kódovanie zložitých naratívnych štruktúr alebo interpretáciu nejednoznačných prvkov v zdrojovom texte.

Ďalšou výhodou automatizácie je konzistentnosť. Metadátové polia, ako sú dátumy publikácií, mená autorov a identifikátory projektov, musia byť formátované jednotným spôsobom, aby bolo možné neskôr texty spoľahlivo triediť, filtrovať alebo vyhľadávať. Nekonzistentnosti - ako sú pravopisné varianty, nesprávne umiestnená interpunkcia alebo nejednotné formáty dátumov - môžu znížiť hodnotu súboru údajov a viesť k zmätkom alebo nesprávnej interpretácii. Automatizované procesy presadzujú jednotnosť tým, že vo všetkých dokumentoch uplatňujú rovnaké pravidlá a šablóny.

Automatizácia podporuje aj opakované použitie a interoperabilitu. V mnohých prípadoch už príslušné metadáta existujú v iných štruktúrovaných formách: v knižničných katalógoch, tabuľkách, záznamoch MARC alebo inštitucionálnych databázach. Automatizácia

umožňuje mapovať tieto údaje priamo do TEI štruktúr, čím sa minimalizuje duplicitné úsilie a maximalizuje sa súlad so zavedenými zdrojmi metadát. To je dôležité najmä v projektoch, ktorých cieľom je prepojenie s inými digitálnymi zbierkami, repozitármi alebo vedeckými infraštruktúrami.

Okrem toho automatizácia postupov generovania metadát umožňuje škálovateľnosť a opakovateľnosť ich tvorby. Po vytvorení automatizovaného postupu je možné rýchlo a systematicky pridávať do korpusu nové diela. Ak sa metadátový model projektu zmení, aktualizácie sa môžu spätne aplikovať na všetky záznamy prostredníctvom rovnakých automatizačných skriptov. Táto schopnosť rozširovať a udržiavať metadáta v priebehu času je nevyhnutná pre dlhodobú udržateľnosť projektov.

Napokon automatizácia znižuje riziko ľudskej chyby. Manuálne zadávanie údajov je na ne náchylné, najmä ak sa vykonáva opakovane počas dlhého obdobia. Dokonca aj malé nezrovnalosti - ako napríklad nesprávne vnorenie elementov, nezapísané znaky alebo chýbajúce polia - môžu viesť k nevalidným XML dokumentom alebo znemožniť korektný beh na dátach závislých aplikáciách, akými sú napríklad fazetové vyhľadávacie rozhrania. Automatizované generovanie metadát, ak je správne implementované a otestované, vytvára dobre formulované, predvídateľné a štandardy spĺňajúce výstupy.

Zdroje metadát

Na efektívne vyplnenie TEI záhlavia je potrebné najprv identifikovať a získať relevantné údaje z dôveryhodných zdrojov. Tieto zdroje sa líšia v závislosti od dostupnosti existujúcich informácií, inštitucionálnej infraštruktúry a historického alebo bibliografického charakteru textov. Pochopenie toho, kde a ako tieto informácie nachádzať, je kľúčovým krokom k automatizovanej, ale aj manuálnej tvorbe metadát.

Knižničné katalógy a bibliografické databázy

Pravdepodobne najštrukturovanejší a najautoritatívnejší zdroj metadát pochádza z katalógov národných, akademických alebo verejných knižníc. Tieto systémy - napríklad WorldCat, Európska knižnica alebo národné systémy ako SUDOC (Francúzsko), COPAC (Spojené kráľovstvo) alebo COBISS (Slovinsko) - zvyčajne poskytujú bibliografické záznamy

v štandardizovaných formátoch (napr. MARC21, Dublin Core), ktoré možno konvertovať do XML kompatibilného s TEI. Tieto záznamy často obsahujú cenné údaje, ako napr.:

- Celé meno autora a dátumy narodenia/úmrtia
- Názov a podnázov diela
- Dátum a miesto vydania
- Údaje o vydavateľovi
- ISBN alebo iné identifikátory
- Informácie o jazyku a vydaní

V niektorých prípadoch je možné hromadne exportovať celé kolekcie bibliografických údajov a mapovať ich do príslušných polí TEI záhlavia pomocou konverzných skriptov.

Existujúce digitálne vydania

Ďalším bežným zdrojom metadát sú predchádzajúce digitálne vydania toho istého diela, najmä ak boli publikované v štruktúrovaných formátoch. Platformy digitálnych knižníc, ako sú Project Gutenberg, HathiTrust alebo Internet Archive, často vkladajú metadáta do svojich HTML alebo MARC záznamov. Hoci kvalita a úplnosť týchto informácií sa líši, takéto vydania často ponúkajú dobrý východiskový bod.

Tlačené vydania

Ak neexistujú spoľahlivé digitálne záznamy, ako primárny zdroj metadát slúžia samotné pôvodné tlačené vydania. Kľúčové informácie, ako je úplný názov, meno autora, vydavateľ, dátum a údaj o vydaní, sa zvyčajne dajú prepísať z titulnej strany a iných predlôh (napr. kolofónov, predsádok). Táto metóda je síce prácnejšia, ale zabezpečuje, že metadáta presne odrážajú konkrétne vydanie použité v digitálnom korpuse.

Vedecké bibliografie a literárne databázy

Literárne bibliografie a špecializované databázy (napr. VD18 pre nemecké tlače 18. storočia, Gallica pre francúzske diela alebo Bibliographie de la littérature française) poskytujú kurátorské metadáta zamerané na konkrétne literárne tradície alebo obdobia. Tieto zdroje často obsahujú prvky s pridanou hodnotou, ako je žánrová klasifikácia,

historické poznámky a informácie o prvých vydaniach alebo kritickej recepcii - všetky tieto prvky môžu obohatiť TEI záhlavie nad rámec úplného bibliografického minima.

Inštitucionálne tabuľky a projektová dokumentácia

V kolaboratívnych projektoch sa metadáta často organizujú a udržiavajú v zdieľaných tabuľkách alebo databázach (napr. Google Spreadsheets, CSV súbory, Airtable). Tieto zdroje často slúžia ako plánovacie nástroje a zároveň centrálne úložiská na agregáciu metadát. Ak sú dôsledne formátované, možno ich ľahko importovať do XML dokumentu pomocou nástrojov na automatizovanú transformáciu. Tieto zdroje špecifické pre projekt často obsahujú aj špeciálne informácie vzťahujúce sa samotný proces kódovania, ako napríklad:

- Kto text prepísal alebo zakódoval
- Dátum digitalizácie
- Dodržiavané usmernenia pre kódovanie
- odkazy na zdrojové skeny alebo súbory

Autoritatívne súbory a slovníky

Na zabezpečenie konzistentnosti metadátových záznamov, najmä v prípade názvov a miest, je užitočné nahliadnuť do autoritatívnych súborov, ako sú napr.:

- **VIAF** (Virtual International Authority File) pre mená osôb
- **Geonames** pre geografické entity
- **Wikidata** pre štruktúrované, viacjazyčné údaje o entitách

Metódy a nástroje automatizácie

Skriptovacie jazyky a knižnice

Python a R patria medzi najpoužívanejšie skriptovacie jazyky na automatizáciu extrakcie dát, ich transformácie a vkladania do XML elementov. Python knižnice ako lxml, xml.etree.ElementTree a BeautifulSoup poskytujú výkonné nástroje na parsovanie, navigáciu a úpravu XML dokumentov, takže sú ideálne na prácu s textami kódovanými

v TEI. Okrem toho je knižnica pandas obzvlášť užitočná na čítanie a manipuláciu so štruktúrovanými metadátami uloženými vo formátoch ako CSV alebo Excel, čo umožňuje mapovanie tabuľkových údajov priamo na elementy TEI. Tieto nástroje možno spoločne kombinovať vo vlastných programoch, ktoré automatizujú generovanie kompletných TEI záhlaví

Príklad použitia: Skript v jazyku Python načíta bibliografické metadáta z tabuľky a vygeneruje sadu sekcií TEI záhlavia, pričom vloží hodnoty do preddefinovaných XML šablón.

XSLT transformácie

XSLT (Extensible Stylesheet Language Transformations) je výkonná metóda na transformáciu dokumentov XML z jednej štruktúry do druhej.

- Slúži na konverziu XML z iných XML formátov do TEI kompatibilného s ELTeC.
- Umožňuje reštrukturalizáciu polí metadát alebo normalizáciu hodnôt.
- Vhodný najmä na opakované transformácie vo veľkých korpusoch.

Príklad použitia: Transformácia dokumentov TEI P4 na TEI P5 so štruktúrou v súlade s ELTeC pomocou súboru štýlov.

XML šablóny

Statické súbory XML šablón môžu byť vytvorené so zástupnými znakmi, ktoré môžu byť neskôr dynamicky nahradené hodnotami špecifickými pre projekt alebo konkrétny text. Tieto šablóny môžu byť vyplnené nástrojmi ako Jinja2 (Python), Handlebars.js alebo jednoduchými "search and replace" skriptami.

Príklad použitia: Použitie šablóny s tokenmi `{title}`, `{author}`, `{pubDate}` na vytvorenie stoviek TEI záhlaví s jednotnou štruktúrou.

Generovania TEI záhlaví v Dispre

V tomto projekte bola automatizácia generovania záhlaví dokumentov vytvorená podľa zjednodušeného a modulárneho pracovného postupu s použitím JSON ako ústredného

sprostredkovateľského formátu. Proces sa začína vytvorením štruktúrovaného JSON súboru, do ktorého sa ručne alebo programovo zadajú metadáta pre daný titul - ako napríklad meno autora, názov, dátum vydania, jazyk a zdroj. Táto JSON štruktúra slúži ako čistá, strojovo čitateľná reprezentácia bibliografických a kontextových údajov, ktorú možno ľahko rozšíriť alebo overiť.

Pre ilustráciu uvádzame takýto JSON súbor jedného z diel nachádzajúcich sa v našej zbierke:


```
{
    "language": "slk",
    "editor": "Marek Vician",
    "eltec_edition_str": "ELTeC edícia",
    "srced": {
        "title": "Reštavrácia",
        "publisher": "Slovenský Tatran",
        "pub_place": "Bratislava",
        "pub_date": "1976"
    },
    "firsted": {
        "pub_date": "1860"
    },
    "authors": [
        {
            "first_name": "Ján",
            "last_name": "Kalinčiak",
            "birth_date": "1822",
            "death_date": "1871",
            "ref": "viaf:20476423"
        }
    ],
    "languages": [
        "slo"
    ],
```

```
    "encoding_lvl": "eltec-1",
    "gender": "M",
    "canonicity": "unspecified"
}
```

Po usporiadaní metadát do formátu JSON je ďalším krokom použitie programu Pandoc v kombinácii s vlastnými Lua filtrami na transformáciu týchto údajov do správne štruktúrovaného záhlavia. Pandoc prečíta JSON súbor ako vstupné metadáta a spracuje predpripravenú XML šablónu, ktorá obsahuje zástupné znaky pre rôzne polia záhlavia. Lua filtre sú zodpovedné za mapovanie JSON kľúčov na zodpovedajúce TEI elementy a zabezpečujú, aby sa každá informácia správne vložila do TEI štruktúry.

Výsledkom je potom nasledujúci XML fragment:

```
<tei:teiHeader>
 <tei:fileDesc>
  <tei:titleStmt>
   <tei:title>
    Márnosť všetkého : ELTeC edícia
   </tei:title>
   <tei:author ref="viaf:14806974">
    Slančíkova-Timrava, Božena (1867 - 1951)
   </tei:author>
   <tei:respStmt>
    <tei:resp>
     editor
    </tei:resp>
    <tei:name>
     Marek Vician
    </tei:name>
   </tei:respStmt>
  </tei:titleStmt>
  <tei:extent>
   <tei:measure unit="words">
    10911
```

```
  </tei:measure>
 </tei:extent>
 <tei:publicationStmt>
  <tei:p/>
 </tei:publicationStmt>
 <tei:sourceDesc>
  <tei:bibl type="printSource">
   <tei:author>
    Slančíkova-Timrava, Božena
   </tei:author>
   <tei:title>
    Márnosť všetkého
   </tei:title>
   <tei:pubPlace>
    Turčiansky Sv. Martin
   </tei:pubPlace>
   <tei:publisher>
    Kníhtlačiarsky účastinársky spolok
   </tei:publisher>
   <tei:date>
    1908
   </tei:date>
  </tei:bibl>
 </tei:sourceDesc>
</tei:fileDesc>
<tei:encodingDesc n="eltec-1">
 <tei:p/>
</tei:encodingDesc>
<tei:profileDesc xmlns:eltec="http://distantreading.net/eltec/ns">
 <tei:langUsage>
  <tei:language ident="slk"/>
 </tei:langUsage>
 <tei:textDesc>
```

Kódovanie: štruktúrovanie textu podľa schémy ELTeC

```
    <eltec:authorGender key="M"/>
    <eltec:size key="short"/>
    <eltec:canonicity key="unspecified"/>
    <eltec:timeSlot key="T4"/>
   </tei:textDesc>
  </tei:profileDesc>
  <tei:revisionDesc>
   <tei:change when="2025-02-10">
    Initially created as an ELTeC file
   </tei:change>
  </tei:revisionDesc>
 </tei:teiHeader>
```

Táto metóda ponúka jasné oddelenie medzi obsahom (v JSON) a výstupným formátom (v TEI), čo nielen uľahčuje automatizáciu, ale aj zvyšuje konzistenciu v celom korpuse. Okrem toho, keďže JSON je ľahko čitateľný a editovateľný, poskytuje pohodlný vstupný bod pre používateľov menej oboznámených s XML formátom, zatiaľ čo použitie Pandocu a programovacieho jazyka Lua zabezpečuje, že konečný výstup je v súlade so špecifikáciou ELTeC.

Kódovanie tela textu

Po príprave metadát pre ich vloženie do záhlavia dokumentu je ďalším krokom kódovanie vlastného textu literárnych diel. Tento proces zahŕňa štruktúrovanie obsahu každého diela spôsobom, ktorý nielen zachytáva jeho lineárne rozprávanie, ale odráža aj jeho vnútorné členenie, vlastnosti a interpretačný potenciál.

Kódovanie tela textu si vyžaduje starostlivú pozornosť venovanú logickým a sémantickým štruktúram textu. To zahŕňa identifikáciu a označovanie segmentov, ako sú kapitoly, odseky, nadpisy, hovoriaci (v dialógoch), veršované riadky (ak je to žiadúce) a iné významové členenie.

V našom projekte sme pre kódovanie obsahu diel vytvorili vlastnú poloautomatizovanú procedúru. Namiesto okamžitej práce so surovým textom alebo vytvárania komplexných

parserov na mieru sa v pracovnom postupe zaviedol ľahký medzistupeň, v ktorom bol zdrojový text najprv pripravený a anotovaný v dokumente Microsoft Word (.docx) s použitím jeho známych formátovacích nástrojov na kódovanie štrukturálnych prvkov. Táto metóda využila prístupnosť textových procesorov pre editorov a zároveň si zachovala vysoký stupeň kontroly nad štrukturálnym značením. Formát .docx, interne založený tiež na XML, slúži vstup pre následnú konverziu pomocou programu Pandoc, rozšíreného o náš vlastný Lua filter určený pre generovanie validných ELTeC dokumentov.

Anotovanie štruktúry v programe Microsoft Word

Anotácia surových textov spočívala v aplikácii štýlov a formátovania programu Microsoft Word, ktoré sa štandardne používajú na vizuálnu a sémantickú identifikáciu častí dokumentu, ako sú kapitoly, odseky, nadpisy a iné textové členenia. Táto fáza funguje ako most medzi surovým textom a strojovo čitateľným značkovaním a je prístupná najmä pre redaktorov alebo spolupracovníkov bez rozsiahleho technického zázemia.

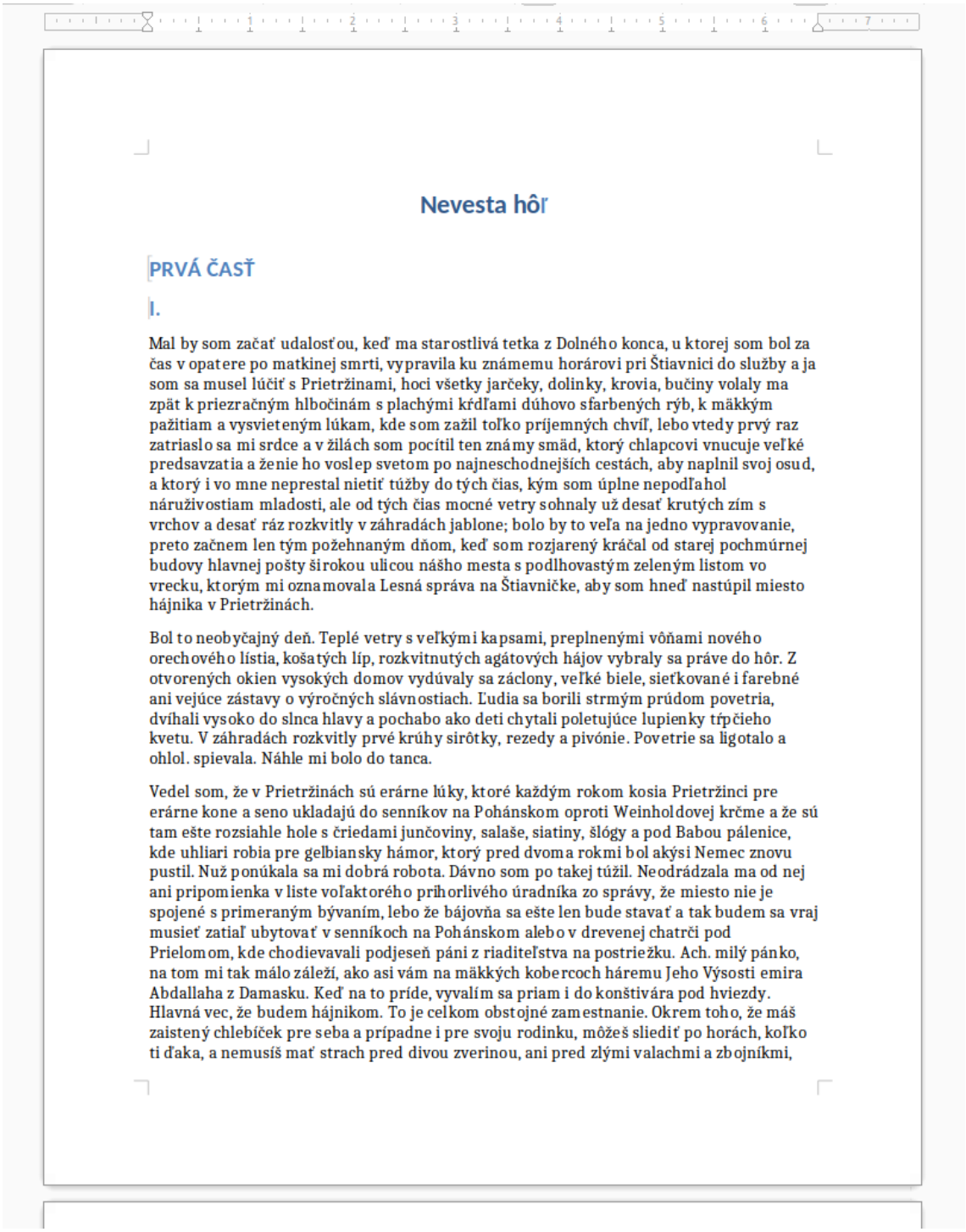
Transformácia anotovaného .docx súboru do XML

Po dokončení štrukturálnej anotácie textu v programe Microsoft Word sa dokument uložený vo formáte .docx dostane do ďalšej fázy pracovného postupu: automatizovanej transformácie do XML súboru. Táto transformácia sa vykonáva opäť pomocou programu Pandoc rozšíreného o náš vlastný Lua filter.

Pandoc je na túto úlohu ideálny vďaka svojej flexibilite pri spracovaní rôznych vstupných formátov (vrátane formátu .docx) a schopnosti generovania štruktúrovaného XML. Pandoc však sám o sebe nepodporuje úplný model TEI - tu zohráva rozhodujúcu úlohu filter napísaný v programovacom jazyku Lua.

Lua filter rozširuje interný model dokumentu Pandoc tak, že reprezentuje štýly a štruktúry .docx v zmysle sémantiky TEI-XML. Napríklad:

- Nadpis v programe Word je mapovaný na elementy `<div type="chapter">` a `<head>` so zodpovedajúcim obsahom
- Z odsekov sa stávajú elementy `<p>`
- Zvýraznenie (napr. kurzíva) sa vykresľuje ako `<hi rend="italic">`.



Nevesta hôľ

PRVÁ ČASŤ

I.

Mal by som začať udalosťou, keď ma starostlivá tetka z Dolného konca, u ktorej som bol za čas v opatere po matkinej smrti, vypravila ku známemu horárovi pri Štiavnici do služby a ja som sa musel lúčiť s Prietržinami, hoci všetky jarčeky, dolinky, krovia, bučiny volaly ma zpäť k priezračným hlbočinám s plachými kŕdľami dúhovo sfarbených rýb, k mäkkým pažitiam a vysvieteným lúkam, kde som zažil toľko príjemných chvíľ, lebo vtedy prvý raz zatriaslo sa mi srdce a v žilách som pocítil ten známy smäd, ktorý chlapcovi vnucuje veľké predsavzatia a ženie ho voslep svetom po najneschodnejších cestách, aby naplnil svoj osud, a ktorý i vo mne neprestal nietiť túžby do tých čias, kým som úplne nepodľahol náruživostiam mladosti, ale od tých čias mocné vetry sohnaly už desať krutých zím s vrchov a desať ráz rozkvitly v záhradách jablone; bolo by to veľa na jedno vypravovanie, preto začnem len tým požehnaným dňom, keď som rozjarený kráčal od starej pochmúrnej budovy hlavnej pošty širokou ulicou nášho mesta s podlhovastým zeleným listom vo vrecku, ktorým mi oznamovala Lesná správa na Štiavničke, aby som hneď nastúpil miesto hájnika v Prietržinách.

Bol to neobyčajný deň. Teplé vetry s veľkými kapsami, preplnenými vôňami nového orechového lístia, košatých líp, rozkvitnutých agátových hájov vybraly sa práve do hôr. Z otvorených okien vysokých domov vydúvaly sa záclony, veľké biele, sieťkované i farebné ani vejúce zástavy o výročných slávnostiach. Ľudia sa borili strmým prúdom povetria, dvíhali vysoko do slnca hlavy a pochabo ako deti chytali poletujúce lupienky tŕpčieho kvetu. V záhradách rozkvitly prvé krúhy sirôtky, rezedy a pivónie. Povetrie sa ligotalo a ohlol. spievala. Náhle mi bolo do tanca.

Vedel som, že v Prietržinách sú erárne lúky, ktoré každým rokom kosia Prietržinci pre erárne kone a seno ukladajú do senníkov na Pohánskom oproti Weinholdovej krčme a že sú tam ešte rozsiahle hole s čriedami jančoviny, salaše, siatiny, šlógy a pod Babou pálenice, kde uhliari robia pre gelbiansky hámor, ktorý pred dvoma rokmi bol akýsi Nemec znovu pustil. Nuž ponúkala sa mi dobrá robota. Dávno som po takej túžil. Neodrádzala ma od nej ani pripomienka v liste voľaktorého prihorlivého úradníka zo správy, že miesto nie je spojené s primeraným bývaním, lebo že bájovňa sa ešte len bude stavať a tak budem sa vraj musieť zatiaľ ubytovať v senníkoch na Pohánskom alebo v drevenej chatrči pod Prielomom, kde chodievavali podjeseň páni z riaditeľstva na postriežku. Ach. milý pánko, na tom mi tak málo záleží, ako asi vám na mäkkých kobercoch háremu Jeho Výsosti emira Abdallaha z Damasku. Keď na to príde, vyvalím sa priam i do konštivára pod hviezdy. Hlavná vec, že budem hájnikom. To je celkom obstojné zamestnanie. Okrem toho, že máš zaistený chlebíček pre seba a prípadne i pre svoju rodinku, môžeš sliediť po horách, koľko ti ďaka, a nemusíš mať strach pred divou zverinou, ani pred zlými valachmi a zbojníkmi,

Obrázok 1: Dokument anotovaný prostriedkami MS Word

Výsledkom transformácie je správne utvorený XML súbor, ktorý je validný vzhľadom na základné štrukturálne požiadavky kladené na dokumenty štandardami ELTeC-u.

Načrtnutý postup ponúka niekoľko kľúčových výhod, ktoré zlepšujú technické aj redakčné aspekty tvorby digitálnej zbierky. Po prvé, výrazne zvyšuje efektívnosť: po správnom naštýlovaní dokumentu pomocou vstavaných formátovacích nástrojov programu Word je proces transformácie rýchly a opakovateľný, čo znižuje potrebu manuálnych zásahov. Po druhé, podporuje konzistentnosť v celom korpuse, pretože Lua filter aplikovaný počas transformácie zabezpečuje, že rovnaké štrukturálne prvky sú jednotne mapované vo všetkých textoch.

Ďalšou dôležitou výhodou je užívateľská prívetivosť. Editori môžu vykonávať všetky štrukturálne anotácie priamo v programe Word bez toho, aby museli pracovať s relatívne zložitou syntaxou XML. A napokon, modularita pracovného postupu umožňuje flexibilitu ďalšieho spracovania; výsledné XML súbory možno obohatiť, validovať alebo ďalej transformovať pomocou ďalších nástrojov kompatibilných s TEI alebo prostredníctvom vlastných skriptov, čo podporuje škálovateľný a udržateľný redakčný proces.

Prezentácia: Publikovanie pomocou TEI Publisher

Po digitalizácii a zakódovaní textov zbierky v súlade s usmerneniami ELTeC a TEI je ďalším krokom ich sprístupnenie na webe. Pri projekte DISPRO sme za týmto účelom využili platformu TEI Publisher: robustný a rozšíriteľný systém určený na transformáciu TEI dokumentov na dynamické, interaktívne webové publikácie.

TEI Publisher vypĺňa medzeru medzi komplexným TEI obsahom a intuitívnym, používateľsky prívetivým digitálnym rozhraním. Jeho použitie nepredpokladá hlboké znalosti programovania alebo webových technológií a ponúka nízkoprahové, ale výkonné prostredie, vďaka ktorému môžu editori a projektové tímy sprístupniť a prezentovať svoje zbierky širokej verejnosti. Prostredníctvom automatizovaných vykresľovacích kanálov, prispôsobiteľných šablón a integrovaných funkcií vyhľadávania a prehliadania umožňuje TEI Publisher vystavovať zakódované texty online pri zachovaní ich základných detailov a štruktúry.

V tejto kapitole sa oboznámime s princípmi, na ktorých TEI Publisher funguje, s technológiami, na ktorých je postavený, a s tým, ako bol použitý v našom projekte. Preskúmame, ako sa dokumenty indexujú a zobrazujú, ako sa konfigurujú používateľské rozhrania a ako sa implementujú funkcie, ako je fazetové vyhľadávanie alebo porovnávanie verzií. Ukážeme pritom, ako môže TEI Publisher slúžiť ako publikačný základ pre digitálne literárne edície a zbierky.

Základné princípy TEI Publisher-a

V jadre sa TEI Publisher riadi súborom zásad, ktoré odrážajú hodnoty digitálnej vedy a praktické potreby publikovania štruktúrovaných textových údajov. Jeho dizajnová filozofia kladie dôraz na transparentnosť, udržateľnosť, interoperabilitu a použiteľnosť.

Oddelenie obsahu a prezentácie

TEI Publisher striktne oddeľuje zakódovaný obsah (t.j., dokumenty vo formáte TEI XML) a jeho prezentáciou na webe. To zabezpečuje, že textové údaje zostanú čisté, sémanticky transparentné a opakovane použiteľné, kým ich vzhľad a logiku interakcie možno nezávisle upravovať prostredníctvom šablón a konfiguračných súborov. Takáto modularita podporuje dlhodobé uchovávanie aj flexibilné publikačné stratégie.

Štandardizácia

Namiesto vyvíjania proprietárneho systému vychádza TEI Publisher z otvorených štandardov a všeobecne prijatých technológií. Základom je samotný formát TEI XML, ale platforma využíva aj XSLT na transformácie, XQuery a XPath na vyhľadávanie a indexovanie a webové technológie, ako sú HTML, CSS a JavaScript, na prezentáciu v užívateľských rozhraniach. Toto opieranie sa o štandardy podporuje kompatibilitu so širším technologickým ekosystémom a dlhú životnosť, bez vytvárania závislosti od výrobcu aplikácie.

Deklaratívna konfigurácia

TEI Publisher uprednostňuje deklaratívny prístup pred pevne zakódovanou logikou. Editori a vývojári definujú, ako sa má obsah zobrazovať alebo vyhľadávať pomocou konfiguračných XML súborov, úprav ODD a transformačných šablón, bez potreby písania zložitého kódu. Tým sa znižujú technické prekážky a humanisti majú možnosť priamo formovať rozhranie a správanie svojich edícií.

Modelmi riadené transformácie

Dôležitou zásadou je, že publikovanie by sa malo riadiť logikou samotného dátového modelu. Štrukturálne a sémantické informácie obsiahnuté v zdrojových dokumentoch tak riadia spôsob vykresľovania textov a interakciu s nimi na webe.

Okamžitá spätná väzba a itertívny vývoj

Keďže TEI Publisher automatizuje veľkú časť procesu transformácie, podporuje tiež rýchly a reprodukovateľný publikačný postup. Redaktori si môžu prezerať zmeny súborov štýlov, šablón alebo obsahu dokumentov s minimálnym časovým odstupom, čo podporuje experimentovanie a dolaďovanie bez zdĺhavých krokov kompilácie alebo nasadenia.

Užívateľská prívetivosť

TEI Publisher je navrhnutý s ohľadom na prístupnosť pre širokú škálu používateľov - od vedcov s malými technickými skúsenosťami až po vývojárov, ktorí vytvárajú vlastné rozhrania. Jeho používateľsky prívetivý ovládací panel, dokumentácia a prispôsobiteľnosť transformačných dokumentov (ODD) uľahčujú spoluprácu aj vo veľkých projektových tímoch.

Systémové nastavenie pre beh TEI Publisher-a

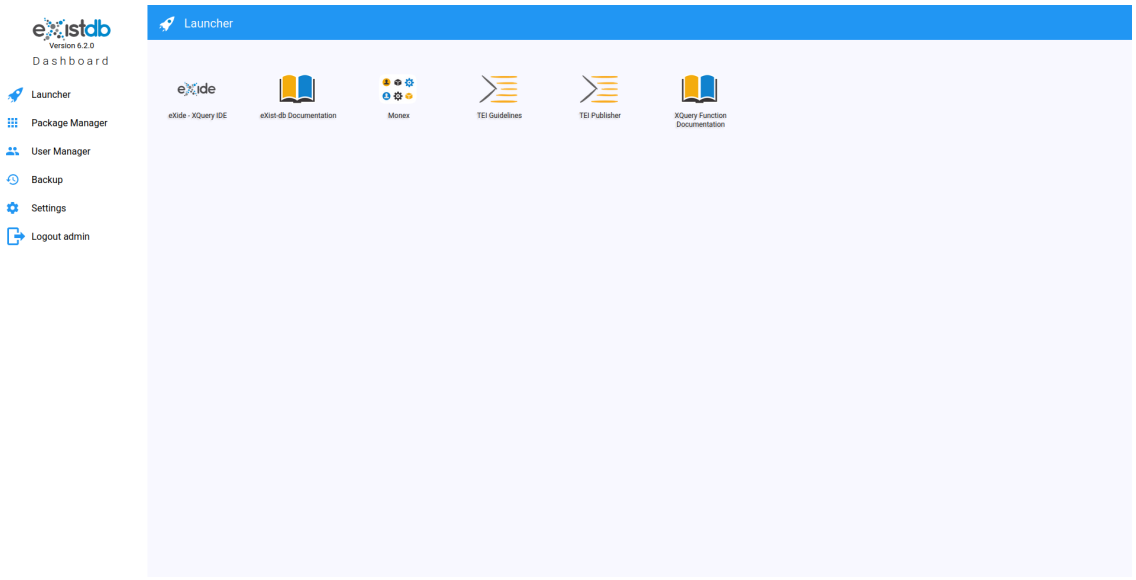
Pred použitím programu je nevyhnutné zvoliť si a pripraviť systémové prostredie, v ktorom bude operovať. TEI Publisher je navrhnutý ako multiplatformová a užívateľsky prívetivá aplikácia, ale spolieha sa na niekoľko kľúčových technológií, ktoré musia byť správne nainštalované a nakonfigurované. V tejto časti uvádzame kroky a komponenty potrebné pre funkčné nastavenie.

Výber hostiteľského prostredia

TEI Publisher je postavený na eXist-db, open-source natívnej XML databáze, ktorá podporuje XQuery, XSLT a XPath. Prvým krokom je rozhodnutie, kde bude táto databáza nasadená:

- Lokálna inštalácia: vhodné na vývoj a experimentovanie.
- Inštitucionálny server: vhodné pre produkčné prostredia s verejným prístupom.
- Cloudový hosting: vhodné pre projekty, ktoré nemajú inštitucionálny server.

Inštalácia eXist-db



Obrázok 2: Administratívny panel eXist-d

Základnou požiadavkou je funkčná inštalácia eXist-db (odporúča sa verzia 6.0 alebo novšia). Inštalačné balíky sú k dispozícii pre systémy Windows, macOS a Linux. Po inštalácii musí byť spustený databázový server a panel správcu by mal byť prístupný prostredníctvom prehliadača buď na adrese `http://localhost:8080`, ak ide o lokálnu inštaláciu, alebo na verejnej adrese pridelenej nášmu serveru.

Nasadenie TEI Publisher-a

Po spustení eXist-db:

1. Treba stiahnuť .xar balík aplikácie TEI Publisher z oficiálneho GitHub repozitára.
2. Nahrať a nainštalovať balík pomocou správcu balíkov na paneli eXist-db.
3. Spustiť aplikáciu na konkrétnom koncovom bode (napr. `http:localhost:8080/tei-publisher`, ak pracujeme s lokálnou inštaláciou).

Adresárová štruktúra projektu

Projekty sú zvyčajne uložené v určitom adresári aplikácie v prostredí eXist-db. Typická štruktúra podadresárov zahŕňa:

- data/ - obsahuje TEI XML súbory.
- resources/ - CSS, JS, obrázky a iné zdroje.
- templates/ - XSLT šablóny alebo HTML na vykresľovanie obsahu.
- config.xml a odd.xml - konfiguračné a prispôsobovacie súbory projektu.

Konfigurácia projektu

Po nasadení je potrebné nastaviť počiatočnú konfiguráciu:

- cesty k zbierke pre dokumenty TEI
- Predvolené šablóny pre vykresľovanie prvkov
- dotazy na navigáciu a prehliadanie
- pravidlá zobrazovania metaúdajov

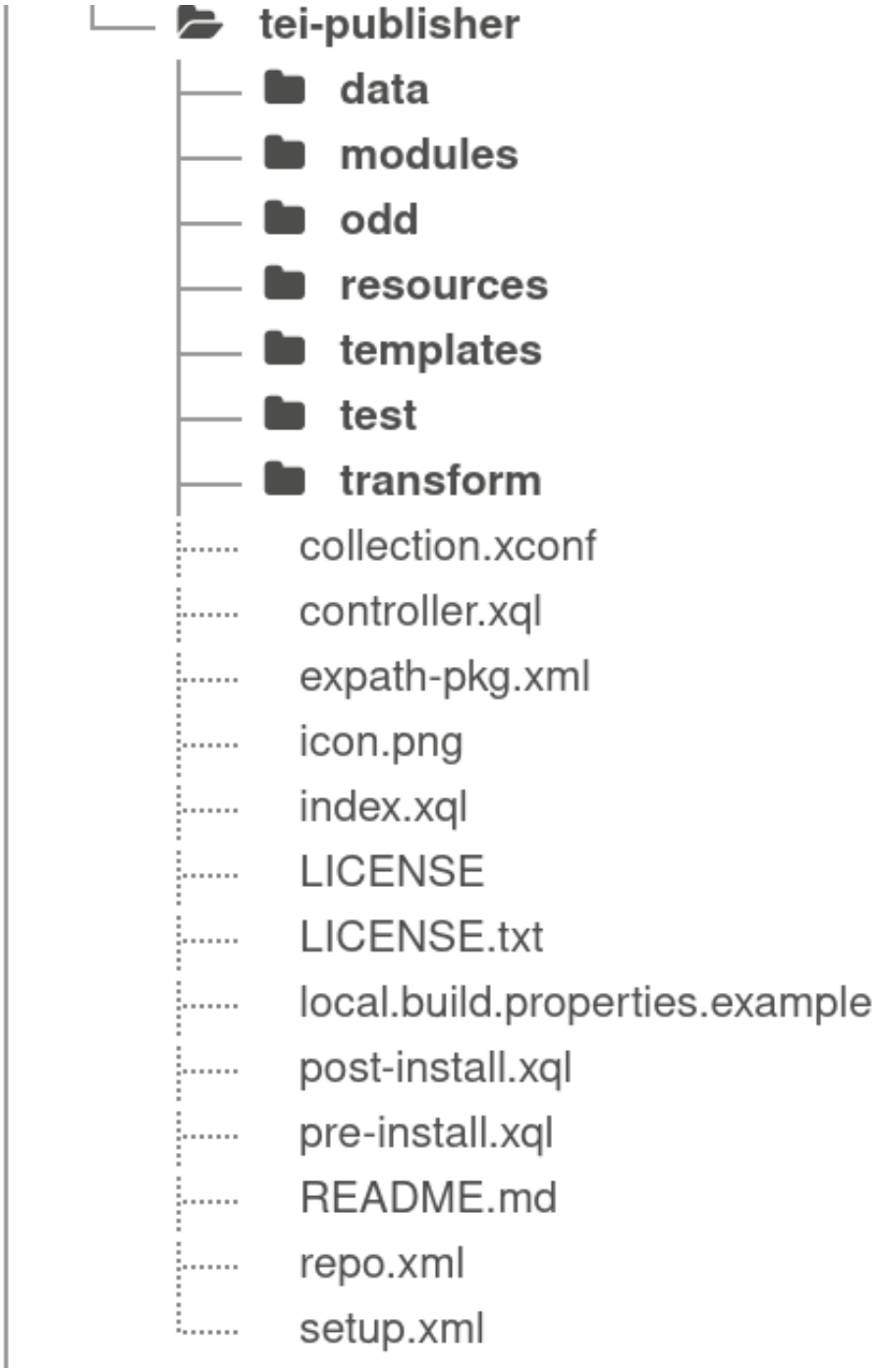
čo sa zvyčajne vykonáva v súbore config.xml a/alebo prostredníctvom ovládacieho panela.

Načítanie dokumentov

Importovanie TEI dokumentov adresára `data/` pomocou ovládacieho panela, WebDAV protokolu alebo priameho nahrania súborov do daného adresára na serveri. TEI Publisher automaticky rozpozná štrukturálne TEI elementy a umožní základné funkcie prehliadania a zobrazovania dokumentov.

Náhľad a iterácia

Po načítaní dokumentov môžeme použiť vstavaný náhľad na kontrolu zobrazenia kolekcie. Následne môžeme iteratívne upravovať súbory štýlov, šablón a dotazov, za účelom doladenia vzhľadu, navigácie a správania kolekcie.



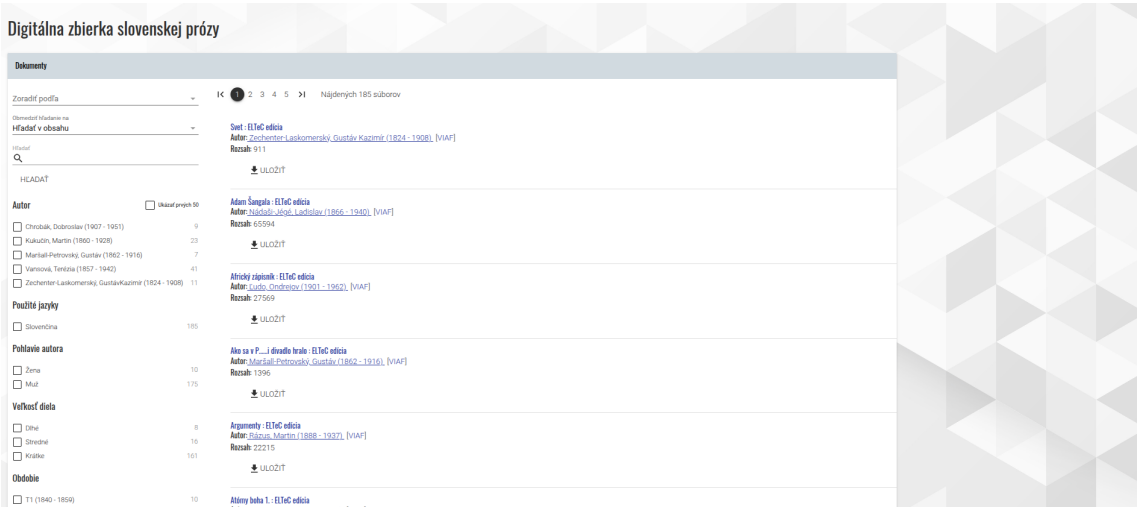
Obrázok 3: Adresárová štruktúra TEI Publisher projektu

Prispôsobenie rozhrania a správania publikovaného webu

Po nainštalovaní TEI Publisher-a a jeho naplnení dokumentmi môžeme pristúpiť k prispôsobeniu používateľského rozhrania a interakcie potrebám konkrétneho projektu alebo publika.

Upravovanie rozvrhnutia a štýlov používateľského rozhrania

TEI Publisher ovláda vykresľovanie obsahu kombináciou šablón, štýlov a prostredníctvom mechanizmov vizuálnych tém. Transformácia prvkov TEI do HTML sa riadi prostredníctvom súborov XSLT umiestnených v projektovom adresári `/templates/`. Tieto šablóny určujú, ako sa štrukturálne a inline prvky interpretujú a zobrazujú na webe. Vizuálna prezentácia sa upravuje pomocou CSS pravidiel, definovaných v súboroch umiestnených v adresári `/resources/css/`. Tie definujú rozloženie, fonty, farebné schémy a responzivitu, čím zabezpečujú, že obsah zostane prístupný a vizuálne koherentný na všetkých zariadeniach. Okrem toho aplikácia podporuje mechanizmus vizuálnych tém, čo umožňuje vývojárom vytvárať a prepínať medzi rôznymi režimami zobrazenia - napríklad tmavým alebo svetlým režimom alebo možnosťami s vysokým kontrastom - prispôsobenými preferenciám používateľov alebo potrebám projektu.



Obrázok 4: Hlavné užívateľské rozhranie Dispra

Obrázok 5: Filtre projektu Dispro

Konfigurácia navigácie a štruktúry dokumentu

TEI Publisher ponúka niekoľko prispôsobiteľných funkcií, ktoré zlepšujú navigáciu a interakciu s digitálnou zbierkou. Obsah sa automaticky generuje zo štrukturálnych prvkov dokumentu - ako napríklad `<div type="chapter">` a `<head>` - ale možno ho ďalej prispôsobiť pomocou vlastných dotazov alebo pravidiel zobrazovania definovaných v súbore `config.xml` umiestnenom v koreňovom adresáre projektu. Fazetové vyhľadávanie využíva metadáta uložené v `<teiHeader>` na vytvorenie dynamických filtrov, ktoré umožňujú používateľom preskúmať zbierku na základe kritérií, ako je autor, dátum alebo žáner. Okrem toho, možno pomocou výrazov XPath konfigurovať "omrvinkovú" navigáciu a názvy stránok, ktoré poskytujú jasnú orientáciu tým, že odrážajú vnútornú hierarchiu dokumentov.

Konfigurácia vyhľadávania

TEI Publisher podporuje fulltextové vyhľadávanie a fazetové dotazy takpovediac "out of box". Vlastné funkcie vyhľadávania je možné vytvoriť:

- Definovaním vlastných vyhľadávacích modulov XQuery.
- Úpravou nastavení indexov v eXist-db s cieľom uprednostniť alebo vylúčiť určité prvky.
- Zvýraznenie výsledkov vyhľadávania v rámci zobrazenia textu.

Rozšírenia dynamických schopností užívateľského rozhrania pomocou jazyka JavaScript

Interaktívne funkcie - ako napríklad prepínanie poznámok pod čiarou, zobrazovanie vyskakovacích okien pre poznámky alebo dynamické načítanie obsahu - sú podporované jazykom JavaScript. Prostredníctvom neho je možné:

- Rozšíriť existujúce komponenty pomocou štandardného JavaScriptu alebo knižníc, ako jQuery.
- Pridať nové správanie používateľského rozhrania prostredníctvom priečinka /resources/scripts/.

- Vytvárať vlastné zobrazenia, napr. pre časové osi, vizualizácie alebo mapy, pomocou externých API rozhraní.

Prispôsobenie vykresľovania TEI elementov

Nie všetky projekty potrebujú rovnaké zaobchádzanie s TEI elementami. Predvolené pravidlá vykresľovania je možné prepísať alebo rozšíriť:

- úpravou šablón XSLT, ktoré definujú ako sa majú elementy zobrazovať alebo správať.
- Pomocou prispôsobenia `odd.xml` definovať, ktoré prvky a atribúty TEI sa použijú a ako sa interpretujú.

Napríklad:

- `<note>` môže byť zobrazený ako vyskakovacie okno, anotácia v bočnom paneli alebo ako riadkový text.
- Element `<foreign>` môže byť zobrazený kurzívou alebo doplnený nápovedou.

Podpora viacjazyčného rozhrania a obsahu

TEI Publisher podporuje internacionalizáciu (i18n) prostredníctvom jazykových súborov, prostredníctvom ktorých je možné lokalizovať reťazce užívateľského rozhrania, alebo umožniť používateľom prepínať medzi jazykmi obsahu, ak sú vaše dokumenty TEI viacjazyčné alebo obsahujú preklady (`<div type="translation">`).

V projekte DISPRO bolo celé užívateľské rozhranie preložené do slovenského jazyka.

Používateľské roly a spravovanie oprávnení

V prostredí s viacerými používateľmi alebo v kolaboratívnych projektoch možno TEI Publisher nakonfigurovať s vrstvami overovania používateľov a oprávnení pomocou bezpečnostných funkcií eXist-db. To umožňuje napríklad udeľovanie práv na úpravu pre autorizovaných prispievateľov, obmedzovanie prístupu verejnosti k draftom alebo citlivým materiálom a podporu vlastných redakčných pracovných postupov prispôsobených potrebám projektu.

Bibliografia

Algee-Hewitt, Mark A., and Mark McGurl. *Between Canon and Corpus: Six Perspectives on 20th-Century Novels.* n.d. Accessed July 24, 2025. https://www.academia.edu/10034192/Between_Canon_and_Corpus_Six_Perspectives_on_20th_Century_Novels.

Jockers, Matthew L. *Macroanalysis: Digital Methods and Literary History.* University of Illinois Press, 2013. https://doi.org/10.5406/illinois/9780252037528.001.0001.

Debnár, Marek - Yesypenko, Dmytro. "Budovanie komplexných a reprezentatívnych digitálnych literárnych zbierok v rámci Európskej zbierky literárnych textov (ELTeC) na Slovensku a Ukrajine." *Slovenská literatúra*, č. 67 (2020): 630–638.

Debnár, Marek – Gogora, Andrej. *Digitálne trendy v súčasných humanitných vedách.* Nitra: UKF, 2019.

Debnár, Marek. "Čítanie z druhej ruky." *World Literature Studies*, č. 3 (2017): 87–97.

Damrosh, David. *What Is World Literature?.* Princeton: Princeton University Press, 2013.

Moretti, Franco. "Conjectures on World Literature." *New Left Review*, no. 1 (February 2000): 54–68.

Moretti, Franco. *Distant Reading.* Verso Books, 2013.

Moretti, Franco. *Graphs, Maps, Trees: Abstract Models for a Literary History*. Verso, 2005.

Nünning, Ansgar. *Lexikon Teorie Literatury a Kultury*. Host, 2025. https://library.upol.cz/arl-upol/en/detail-upol_us_cat-m0156122-Lexikon-teorie-literatury-a-kultury/.

"Tesseract User Manual." In *Tessdoc*. n.d. Accessed July 17, 2025. https://tesseract-ocr.github.io/tessdoc/Home.html.

Transkribus - Unlocking the Past with AI. n.d. Accessed July 17, 2025. https://www.transkribus.org/.

Budovanie digitálnych zbierok 2. (Teória a aplikácia)

Marek Debnár - Marek Vician

Vydavateľ: Univerzita Konštantína Filozofa v Nitre

Prvé vydanie, Nitra 2025

Počet strán: 72

ISBN 978-80-558-2290-7