Dispro učebnica

Obsah

# Úvod do digitálnych humanitných vied

## Definícia DH

Aliquam erat volutpat. Nunc eleifend leo vitae magna. In id erat non orci commodo lobortis. Proin neque massa, cursus ut, gravida ut, lobortis eget, lacus. Sed diam. Praesent fermentum tempor tellus. Nullam tempus. Mauris ac felis vel velit tristique imperdiet. Donec at pede. Etiam vel neque nec dui dignissim bibendum. Vivamus id enim. Phasellus neque orci, porta a, aliquet quis, semper a, massa. Phasellus purus. Pellentesque tristique imperdiet tortor. Nam euismod tellus id erat.

Aliquam erat volutpat. Nunc eleifend leo vitae magna. In id erat non orci commodo lobortis. Proin neque massa, cursus ut, gravida ut, lobortis eget, lacus. Sed diam. Praesent fermentum tempor tellus. Nullam tempus. Mauris ac felis vel velit tristique imperdiet. Donec at pede. Etiam vel neque nec dui dignissim bibendum. Vivamus id enim. Phasellus neque orci, porta a, aliquet quis, semper a, massa. Phasellus purus. Pellentesque tristique imperdiet tortor. Nam euismod tellus id erat.

## História DH

Aliquam erat volutpat. Nunc eleifend leo vitae magna. In id erat non orci commodo lobortis. Proin neque massa, cursus ut, gravida ut, lobortis eget, lacus. Sed diam. Praesent fermentum tempor tellus. Nullam tempus. Mauris ac felis vel velit tristique imperdiet. Donec at pede. Etiam vel neque nec dui dignissim bibendum. Vivamus id enim. Phasellus neque orci, porta a, aliquet quis, semper a, massa. Phasellus purus. Pellentesque tristique imperdiet tortor. Nam euismod tellus id erat.

Aliquam erat volutpat. Nunc eleifend leo vitae magna. In id erat non orci commodo lobortis. Proin neque massa, cursus ut, gravida ut, lobortis eget, lacus. Sed diam. Praesent fermentum tempor tellus. Nullam tempus. Mauris ac felis vel velit tristique imperdiet. Donec at pede. Etiam vel neque nec dui dignissim bibendum. Vivamus id enim. Phasellus neque orci, porta a, aliquet quis, semper a, massa. Phasellus purus. Pellentesque tristique imperdiet tortor. Nam euismod tellus id erat.

## Vzťah DH a tradičných humanitných vied

Aliquam erat volutpat. Nunc eleifend leo vitae magna. In id erat non orci commodo lobortis. Proin neque massa, cursus ut, gravida ut, lobortis eget, lacus. Sed diam. Praesent fermentum tempor tellus. Nullam tempus. Mauris ac felis vel velit tristique imperdiet. Donec at pede. Etiam vel neque nec dui dignissim bibendum. Vivamus id enim. Phasellus neque orci, porta a, aliquet quis, semper a, massa. Phasellus purus. Pellentesque tristique imperdiet tortor. Nam euismod tellus id erat.

Aliquam erat volutpat. Nunc eleifend leo vitae magna. In id erat non orci commodo lobortis. Proin neque massa, cursus ut, gravida ut, lobortis eget, lacus. Sed diam. Praesent fermentum tempor tellus. Nullam tempus. Mauris ac felis vel velit tristique imperdiet. Donec at pede. Etiam vel neque nec dui dignissim bibendum. Vivamus id enim. Phasellus neque orci, porta a, aliquet quis, semper a, massa. Phasellus purus. Pellentesque tristique imperdiet tortor. Nam euismod tellus id erat.

# Ústredné teórie a pojmy

Nullam eu ante vel est convallis dignissim. Fusce suscipit, wisi nec facilisis facilisis, est dui fermentum leo, quis tempor ligula erat quis odio. Nunc porta vulputate tellus. Nunc rutrum turpis sed pede. Sed bibendum. Aliquam posuere. Nunc aliquet, augue nec adipiscing interdum, lacus tellus malesuada massa, quis varius mi purus non odio. Pellentesque condimentum, magna ut suscipit hendrerit, ipsum augue ornare nulla, non luctus diam neque sit amet urna. Curabitur vulputate vestibulum lorem. Fusce sagittis, libero non molestie mollis, magna orci ultrices dolor, at vulputate neque nulla lacinia eros. Sed id ligula quis est convallis tempor. Curabitur lacinia pulvinar nibh. Nam a sapien.

## Interdisciplinarita v DH

Aliquam erat volutpat. Nunc eleifend leo vitae magna. In id erat non orci commodo lobortis. Proin neque massa, cursus ut, gravida ut, lobortis eget, lacus. Sed diam. Praesent fermentum tempor tellus. Nullam tempus. Mauris ac felis vel velit tristique imperdiet. Donec at pede. Etiam vel neque nec dui dignissim bibendum. Vivamus id enim. Phasellus neque orci, porta a, aliquet quis, semper a, massa. Phasellus purus. Pellentesque tristique imperdiet tortor. Nam euismod tellus id erat.

Pellentesque dapibus suscipit ligula. Donec posuere augue in quam. Etiam vel tortor sodales tellus ultricies commodo. Suspendisse potenti. Aenean in sem ac leo mollis blandit. Donec neque quam, dignissim in, mollis nec, sagittis eu, wisi. Phasellus lacus. Etiam laoreet quam sed arcu. Phasellus at dui in ligula mollis ultricies. Integer placerat tristique nisl. Praesent augue. Fusce commodo. Vestibulum convallis, lorem a tempus semper, dui dui euismod elit, vitae placerat urna tortor vitae lacus. Nullam libero mauris, consequat quis, varius et, dictum id, arcu. Mauris mollis tincidunt felis. Aliquam feugiat tellus ut neque. Nulla facilisis, risus a rhoncus fermentum, tellus tellus lacinia purus, et dictum nunc justo sit amet elit.

### Humanitné disciplíny (literatúra, história, filozofia, umenie, lingvistika)

Aliquam erat volutpat. Nunc eleifend leo vitae magna. In id erat non orci commodo lobortis. Proin neque massa, cursus ut, gravida ut, lobortis eget, lacus. Sed diam. Praesent fermentum tempor tellus. Nullam tempus. Mauris ac felis vel velit tristique imperdiet. Donec at pede. Etiam vel neque nec dui dignissim bibendum. Vivamus id enim. Phasellus neque orci, porta a, aliquet quis, semper a, massa. Phasellus purus. Pellentesque tristique imperdiet tortor. Nam euismod tellus id erat.

### Výpočtové metódy (programovanie, dátové modelovanie, strojové učenie)

Pellentesque dapibus suscipit ligula. Donec posuere augue in quam. Etiam vel tortor sodales tellus ultricies commodo. Suspendisse potenti. Aenean in sem ac leo mollis blandit. Donec neque quam, dignissim in, mollis nec, sagittis eu, wisi. Phasellus lacus. Etiam laoreet quam sed arcu. Phasellus at dui in ligula mollis ultricies. Integer placerat tristique nisl. Praesent augue. Fusce commodo. Vestibulum convallis, lorem a tempus semper, dui dui euismod elit, vitae placerat urna tortor vitae lacus. Nullam libero mauris, consequat quis, varius et, dictum id, arcu. Mauris mollis tincidunt felis. Aliquam feugiat tellus ut neque. Nulla facilisis, risus a rhoncus fermentum, tellus tellus lacinia purus, et dictum nunc justo sit amet elit.

Pellentesque dapibus suscipit ligula. Donec posuere augue in quam. Etiam vel tortor sodales tellus ultricies commodo. Suspendisse potenti. Aenean in sem ac leo mollis blandit. Donec neque quam, dignissim in, mollis nec, sagittis eu, wisi. Phasellus lacus. Etiam laoreet quam sed arcu. Phasellus at dui in ligula mollis ultricies. Integer placerat tristique nisl. Praesent augue. Fusce commodo. Vestibulum convallis, lorem a tempus semper, dui dui euismod elit, vitae placerat urna tortor vitae lacus. Nullam libero mauris, consequat quis, varius et, dictum id, arcu. Mauris mollis tincidunt felis. Aliquam feugiat tellus ut neque. Nulla facilisis, risus a rhoncus fermentum, tellus tellus lacinia purus, et dictum nunc justo sit amet elit.

Nullam eu ante vel est convallis dignissim. Fusce suscipit, wisi nec facilisis facilisis, est dui fermentum leo, quis tempor ligula erat quis odio. Nunc porta vulputate tellus. Nunc rutrum turpis sed pede. Sed bibendum. Aliquam posuere. Nunc aliquet, augue nec adipiscing interdum, lacus tellus malesuada massa, quis varius mi purus non odio. Pellentesque condimentum, magna ut suscipit hendrerit, ipsum augue ornare nulla, non luctus diam neque sit amet urna. Curabitur vulputate vestibulum lorem. Fusce sagittis, libero non molestie mollis, magna orci ultrices dolor, at vulputate neque nulla lacinia eros. Sed id ligula quis est convallis tempor. Curabitur lacinia pulvinar nibh. Nam a sapien.

### Knižničné a informačné vedy (metadáta, digitálne archivovanie)

Aliquam erat volutpat. Nunc eleifend leo vitae magna. In id erat non orci commodo lobortis. Proin neque massa, cursus ut, gravida ut, lobortis eget, lacus. Sed diam. Praesent fermentum tempor tellus. Nullam tempus. Mauris ac felis vel velit tristique imperdiet. Donec at pede. Etiam vel neque nec dui dignissim bibendum. Vivamus id enim. Phasellus neque orci, porta a, aliquet quis, semper a, massa. Phasellus purus. Pellentesque tristique imperdiet tortor. Nam euismod tellus id erat.

Aliquam erat volutpat. Nunc eleifend leo vitae magna. In id erat non orci commodo lobortis. Proin neque massa, cursus ut, gravida ut, lobortis eget, lacus. Sed diam. Praesent fermentum tempor tellus. Nullam tempus. Mauris ac felis vel velit tristique imperdiet. Donec at pede. Etiam vel neque nec dui dignissim bibendum. Vivamus id enim. Phasellus neque orci, porta a, aliquet quis, semper a, massa. Phasellus purus. Pellentesque tristique imperdiet tortor. Nam euismod tellus id erat.

### Sociálne vedy

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Donec hendrerit tempor tellus. Donec pretium posuere tellus. Proin quam nisl, tincidunt et, mattis eget, convallis nec, purus. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Nulla posuere. Donec vitae dolor. Nullam tristique diam non turpis. Cras placerat accumsan nulla. Nullam rutrum. Nam vestibulum accumsan nisl.

# Digitálna zbierka slovenskej prózy (prípadová štúdia)

V tejto kapitole opisujeme vývoj digitálneho korpusu slovenskej prózy vydanej pred rokom 1950, pričom osobitnú pozornosť v nej venujeme prieniku literárnej vedy a prostriedkov výpočtových technológií. Cieľom projektu bolo zostaviť reprezentatívnu a na výskum pripravenú zbierku beletristickej prózy, ktorá by odrážala formálny, tematický a jazykový vývoj slovenskej literatúry od polovice 19. storočia do začiatku povojnového obdobia. Korpus, ktorý čerpal z rôznych archívnych zdrojov - vrátane Slovenskej národnej knižnice, univerzitných repozitárov, jazykovedného inštitútu SAV a historických vydavateľských záznamov - pozostáva z desiatok XML súborov zodpovedajúcich schéme TEI [[1]](#footnote-30) podporujúcej prístupy blízkeho aj vzdialeného čítania. Tento štruktúrovaný súbor údajov umožňuje celý rad vedeckých výskumov, od štylistickej analýzy a klasfikácie žánrov až po sieťové modelovanie autorských a publikačných kontextov. Namiesto vytvárania uzavretého kánonu projekt poskytuje otvorenú platformu na skúmanie literárnej histórie prostredníctvom počítačových nástrojov, pričom zostáva zakotvený v interpretačných tradíciách humanitných vied.

## Opis výskumného projektu

## Technologické minimum

### XML

V digitálnych humanitných vedách nie je výber formátu na reprezentáciu textových údajov neutrálnym rozhodnutím - určuje, čo môžeme s textom robiť, ako ho interpretujeme a ako ho zdieľame s ostatnými. Medzi voľby, ktorá sú zrejme najbližšie bežnému užívateľovi, patria formáty textových procesorov, ako napríklad .docx programu Microsoft Word alebo .odt súbory používané v OpenOffice. Takéto programy ponúkajú vizuálne orientované prostredie, v ktorom môžu používatelia písať, upravovať a formátovať texty bez potreby špecializovaných technických znalostí. Funkcie ako tučné písmo, kurzíva, poznámky pod čiarou a nadpisy sú vďaka intuitívnemu užívateľskému rozhraniu ľahko použiteľné a spoluprácu s ďalšími ľuďmi zjednodušujú integrované funkcie komentovania alebo systémy sledovania zmien. Pri bežnom narábaní s textom v digitálnom prostredí sú vďaka takejto jednoduchosti používania procesory jasnou voľbou.

Táto voľba však so sebou nesie určité obmedzenia, pre ktoré nie sú textové procesory, resp. súborové formáty, ktorými tieto programy reprezentujú texty, vhodné pre ciele, ktoré sledujeme v digitálnych humanitných vedách. Tieto obmedzenia nie sú len technickými prekážkami, ale ovplyvňujú aj spôsob interpretácie, zdieľania a uchovávania textov vo vedeckej práci, ktorá čoraz viac závisí od štruktúrovaných, pre stroje čitateľných údajov.

Jedným z najzásadnejších problémov je, že textové procesory sú navrhnuté s ohľadom na vizuálnu prezentáciu textov, nie vzhľadom na ich sémantickú zrozumiteľnosť. Funkcie formátovania textu, ktoré tieto programy poskytujú, sú zamerané na to, ako text vyzerá pre čitateľa: tučné písmo pre zvýraznenie, kurzíva pre nadpisy, zalomenie riadkov pre odseky. Táto prezentácia však v sebe nenesie žiadne informácie o význame alebo funkcii danej časti textu. Tučným písmom zvýraznený výraz v programe Word môže označovať rečníka v divadelnej hre, postavu v románe alebo nadpis v odbornom článku, čo stroj, bez ďalšej informácie, nemôže vedieť. Táto absencia sémantického značenia veľmi sťažuje extrakciu, analýzu alebo opakované spracovanie textu pomocou výpočtovej techniky. Aj keď je vizuálne formátovanie konzistentné, základná štruktúra súboru je zvyčajne neprehľadná, keďže je uložená ako zazipovaná zbierka binárnych súborov, ktoré je ťažké analyzovať bez špecializovaných nástrojov.

Okrem toho sú súbory textových procesorov často nekonzistentné a idiosynkratické. Používatelia volia rôzne formátovanie v závislosti od osobných zvykov, inštitucionálnych šablón alebo predvolených nastavení softvéru. Jedna vedkyňa môže používať kurzívu pre názvy kníh, iný môže používať úvodzovky. Niektorí môžu ručne vkladať zalomenia riadkov, aby vytvorili dojem medzier, iní sa spoliehajú na štýly. Tieto nezrovnalosti sa v spoločných alebo rozsiahlych projektoch rýchlo hromadia, takže automatizované spracovanie alebo analýza sú bez rozsiahleho čistenia a štandardizácie nespoľahlivé.

Ďalšou nevýhodou je netransparentnosť verziovania zmien súborov textového procesora. Word síce ponúka funkcie ako „sledovanie zmien“, tie však nie sú štandardizované ani prenosné medzi rôznymi platformami a nedajú sa ľahko extrahovať alebo analyzovať v priebehu času.

Z hľadiska uchovávania sú formáty textových procesorov tiež príliš krehké. Keďže sa spoliehajú na proprietárne alebo poloproprietárne formáty, sú náchylné na zastarávanie softvéru alebo zmeny v predvolenom správaní v jeho rôznych verziách. Súbor .docx vytvorený v programe Word 2007 sa nemusí správať rovnako v novších verziách alebo v open-source editoroch, čo môže viesť k strate údajov, posunu formátovania alebo k neželaným zmenám v rozložení textu.

Napokon, pre projekty digitálnych humanitných vied, ktorých cieľom je publikovať alebo prezentovať texty na webe, prepojiť ich s metadátami alebo zabezpečiť ich plnotextovú vyhľadateľnosť a analýzu, sú súbory textového procesora jednoducho nevyhovujúce. Konverzia súborov .docx do štruktúrovaných formátov si zvyčajne vyžaduje buď množstvo manuálnej práce alebo použitie externých nástrojov, akým je napríklad program Pandoc, prípadne vlastné skripty - ani tie nám však nepomôžu, ak nemá pôvodný súbor konzistentnú štruktúru.

Hoci teda formáty textových procesorov vyhovujú potrebám bežného písania a akademického publikovania[[2]](#footnote-32), ich obmedzenia sa naplno prejavia vo vzťahu k požiadavkám digitálnej humanitnej práce - konkrétne s potrebou modelovať, analyzovať a uchovávať texty bohatým a štruktúrovaným spôsobom. Práve tu ponúka XML (eXtensible Markup Language) robustnú alternatívu. Je to jazyk navrhnutý na reprezentáciu informácií v štruktúrovanom, pre človeka a stroj čitateľnom formáte. Pri jeho návrhu sa kládol dôraz najmä na jednoduchosť, všeobecnosť a použiteľnosť v prostredí internetu[[3]](#footnote-33) a vyznačuje sa silnou podporou takmer všektých ľudských jazykov vďaka kompatibilite s Unicode štandardom.[[4]](#footnote-34) Hoci mal jazyk XML pôvodne slúžiť najmä na reprezentáciu dokumentov, v súčasnosti sa extenzívne používa na reprezentáciu ľubovoľných dátových štruktúr,[[5]](#footnote-35) napríklad tých, ktoré sa vyskytujú vo webových službách.[[6]](#footnote-36)

#### Sémantická jasnosť a explicitná štruktúra

Jednou z najvýznamnejších výhod jazyka XML je schopnosť sémantického značkovania. Na rozdiel od súborov textových procesorov, ktoré používajú formátovanie predovšetkým na vizuálnu prezentáciu textu, XML umožňuje explicitné definovanie sémantického významu jednotlivých častí textu. Ak chceme, napríklad, v nejakom texte zaznamenať, že určitý reťazec znakov predstavuje meno autora, prostriedkami XML to dosiahneme tak, že danú pasáž uzavrieme v značke [[7]](#footnote-37), ktorá má vopred definovaný význam.[[8]](#footnote-38) Týmto sa stane rola daného reťazca v dokumente explicitná a jednoznačná.

Táto jasnosť sa zreteľnejšie ukáže pri komplexnejších príkladoch. Historický dokument môže obsahovať vrstvené úrovne citácií, redakčných a autorských poznámok, odkazov alebo marginálií - každý z týchto prvkov možno presne reprezentovať označením pomocou prostriedkov XML. Vďaka tomu tak výskumníci môžu systematicky vyhľadávať prípady konkrétneho hovorcu, sledovať pomenované entity, identifikovať tematické vzory alebo rozlišovať medzi pôvodným textom a redakčnými zásahmi.

XML tak slúži ako nástroj pre formalizované vyjadrenie vedeckej interpretácie. Zviditeľňuje štrukturálne a interpretačné rozhodnutia, ktoré humanisti často nechávajú v tradičnej vedeckej produkcii implicitné. To sa obzvlášť dobre zhoduje s cieľmi tvorby kritických edícií a archívnej práce všeobecne, kde je prvoradá vernosť materiálnemu a intelektuálnemu kontextu.

#### Interoperabilita a znovupoužiteľnosť

Ďalšou kľúčovou výhodou XML je jeho interoperabilita. Keďže je nezávislý od výpočtovej platformy a riadi sa otvorenými štandardmi, súbory tohto formátu možno používať v širokom spektre softvérových prostredí, od databáz a webových aplikácií až po transformačné systémy a nástroje na vizualizáciu údajov.

Súbor vo formáte XML možno napríklad transformovať do HTML formátu určeného na publikovanie na webe, PDF formátu vhodného pre tlač, formátu ePub používaného v elektronických čítačkách alebo dokonca do formátu JSON na integráciu do webových rozhraní. Tieto transformácie sa zvyčajne realizujú pomocou XSLT [[9]](#footnote-40) alebo iných transformačných “potrubí”[[10]](#footnote-41), vďaka čomu môže jeden súbor slúžiť ako zdroj pre generovania množstva rôznych výstupov bez vynakladania duplicitnej práce.

Okrem toho, keďže sa XML sa riadi konzistentnými pravidlami a súbory v tomto formáte môžeme validovať voči vopred definovaným schémam, je ľahké udržiavať texty dobre sformované a vnútorne konzistentné. Toto zabezpečuje opakovateľnú použiteľnosť a zdieľateľnosť korpusov pozostávajúcich z XML súborov - nielen pôvodnými autormi, ale aj inými výskumníkmi a inštitúciami.

#### Strojová čitateľnosť a počítačová analýza

Vďaka hierarchickej a na pravidlách založenej štruktúre XML, poskytujú súbory v tomto formáte ideálny substrát pre dištančné čítanie, stylometriu, sieťovú analýzu, modelovanie tém a ďalšie metódy používané v digitálnych humanitných vedách. Vhodne anotované texty nám napríklad umožňujú ľahko zodpovedať otázky ako koľko ženských postáv hovorí v slovenských románoch z 19. storočia, ako často sa objavujú odkazy na určité miesta alebo ako sa mení štruktúra dialógov v čase. Na tieto typy otázok je takmer nemožné spoľahlivo odpovedať pri použití formátov textového procesora, ktoré nemajú vnútornú štruktúru potrebnú na to, aby boli vhodnými vstupmi pre automatizované spracovanie.

#### Transparentnosť a uchovávanie

Keďže XML je čisto textový formát[[11]](#footnote-44), vyznačuje sa transparentnosťou a trvácnosťou. Na rozdiel od proprietárnych formátov textových procesorov možno súbory v tomto formáte otvoriť a čítať v akomkoľvek textovom editore, v akomkoľvek operačnom systéme, bez špeciálneho softvéru.

Vďaka tomuto zmeny v súboroch XML dajú presne sledovať pomocou systémov na kontrolu verzií, ako je napríklad Git, čo je obzvlášť užitočné v kolaboratívnom vedeckom prostredí. Každá úprava, doplnenie alebo oprava sa stáva súčasťou kontrolovateľnej histórie, čo napríklad umožňuje budúcim výskumníkom pochopiť vývoj digitálneho objektu.

To, že XML je čisto textový formát, znamená oddelenie obsahu od prezentácie, čo podporuje čistejšie pracovné postupy a znižuje riziko poškodenia údajov v dôsledku problémov s formátovaním. Prezentácia - či už pre web, tlač alebo mobilné zariadenia - sa dá spracovať nezávisle prostredníctvom súborov štýlov a šablón, pričom základné údaje zostanú nedotknuté.

#### Komunita a štandardy

XML v digitálnych humanitných vedách ťaží zo silných komunít, najmä okolo TEI (Text Encoding Initiative), ktorá poskytuje dobre vyvinutý a vyvíjajúci sa štandard pre textovú vedu. TEI ponúka nielen rozsiahly slovník značiek[[12]](#footnote-46) pre širokú škálu textových funkcií - poskytuje aj dokumentáciu, príklady, nástroje a komunitu vedcov, editorov a vývojárov, ktorí aktívne podporujú jeho prijatie.

Prijatím XML a TEI sa výskumníci zapájajú do ekosystému, ktorý si cení transparentnosť, udržateľnosť a vedeckú prísnosť. Toto zosúladenie so spoločnými štandardmi zvyšuje hodnotu, udržateľnosť a prístupnosť vlastnej práce, čo uľahčuje jej zdieľanie, uchovávanie a budovanie.

#### XML špecifikácia

Pre efektívne používanie XML je dôležité pochopiť jeho základné princípy: pravidlá, ktoré definujú, čo robí dokument XML *správne utvoreným* (“well-formed”) a *validným*. V tejto časti postupne predstavíme kľúčové stavebné prvky tohto formátu, pričom sa technickejšie detaily jeho implementácie budeme snažiť prepájať s abstraktnejšími princípmi, ktorými sme v predchádzajúcom texte motivovali jeho adopciu pre účely digitálnych humanitných vied.

##### Základná štruktúra XML dokumentu

Jadrom každého XML dokumentu je hierachická stromová štruktúra, ktorú nazývame **XML strom**. Tu je príklad minimálneho XML dokumentu:

<?xml version="1.0" encoding="UTF-8" standalone="yes" ?>  
<book>  
 <title>Dom v stráni</title>  
 <author>Martin Kukučín</author>  
 <publisher>Matica Slovenská</publisher>  
 <pubdate>1912</pubdate>  
</book>

Riadok <?xml ... ?> je deklarácia dokumentu, čo je vyhlásenie umiestnené na úplnom začiatku textového súboru, ktoré poskytuje základné informácie o kódovaní[[13]](#footnote-48), verzii[[14]](#footnote-49) a “standalone” stav dokumentu[[15]](#footnote-50). Slúži ako hlavička metadát, ktorá umožňuje parserom a procesorom správne interpretovať obsaho dokumentu.

Element[[16]](#footnote-51) <book> je koreňom dokumentu, pričom každý XML dokument má práve jeden koreňový element. To znamená, že všetky ďalšie elementy, ktoré sa v dokumente nachádzajú, sa nachádzajú vnútri koreňového elementu. Tento hierarchický vzťah sa často metaforicky prezentuje ako vzťah rodiča a potomstva (“parent” - “child”). V predchádzajúcej ilustrácii je teda <book> “rodičom” XML značiek <title>, <author>, <publisher>, <pubdate>, a tie sú zas jeho “potomkovia”. Tento vzťah pritom vyjadruje, že podradený element existuje v konceptuálnej alebo logickej doméne nadradeného prvku.

Pre o niečo komplexnejšiu ilustráciu si vezmime nasledujúcu štruktúru, ktorá určitým spôsobom organizuje informácie o slovenskom spisovateľovi Františkovi Švantnerovi:

<author>  
 <bio>  
 <firstName>František</firstName>  
 <lastName>Švantner</lastName>  
 <birthPlace>Bystrá</birthPlace>  
 <dateOfBirth>29-01-1912</dateOfBirth>  
 <dateOfDeath>13-10-1950</dateOfDeath>  
 </bio>  
 <notableTitles>  
 <title>  
 <name>Nevesta Hôľ</name>  
 <pubDate>1946</pubDate>  
 </title>  
 <title>  
 <name>Život bez konca</name>  
 <pubDate>1946</pubDate>  
 </title>  
 </notableTitles>  
</author>

### XML schéma TEI

### HTML

### CSS

### TEI Publisher

### Linux

Dôležitou, ale často opomínanou zložkou pracovného postupu tvorby korpusu, bolo použitie operačného systému Linux ako technologického základu projektu. Linux poskytoval stabilné prostredie s otvoreným zdrojovým kódom, ktoré sa ideálne hodilo na požiadavky rozsiahleho spracovania textu, kontroly verzií a automatizácie. Jeho kompatibilita so základnými nástrojmi - ako sú knižnice na spracovanie TEI, XML a skriptovacími jazykmi, ako sú Python a Bash - nám umožnil výskumnému tímu vytvoriť vlastné postupy na čistenie údajov, kódovanie a správu korpusu. Okrem toho modulárna konštrukcia systému Linux umožnila jemnú kontrolu nad systémom správania, od oprávnení súborov až po plánovanie úloh, čo sa ukázalo ako nevyhnutné, keď pri práci so súbormi údajov archívneho rozsahu. V tejto časti sa uvádza, ako systém Linux podporoval technickú infraštruktúru projektu, pričom sa zdôrazňuje jeho úloha pri zabezpečovaní transparentnosti, reprodukovateľnosti a dlhodobej udržiavateľnosti - hodnôt, ktoré zdieľa digitálnych humanitných vied a komunitách open-source.

### Programovacie jazyky

Medzi základné technológie použité v projekte budovania korpusu patria programovacie jazyky Python a Lua, ktoré zohrávali odlišné, ale vzájomne sa dopĺňajúce úlohy. Python slúžil ako primárny jazyk na manipuláciu s údajmi, spracovanie textu, a integráciu s knižnicami na spracovanie TEI-XML, parsovanie regulárnych výrazov, a transformáciu metadát. Jeho čitateľnosť, všestrannosť a rozsiahly ekosystém sa hodil na vytváranie robustných skriptov na automatizáciu úloh, ako sú čistenie OCR, overovanie štrukturálnych značiek a štatistické analýzy. Jazyk Lua sme používali predovšetkým na vývoj vlastných filtrov a zapisovačov pre Pandoc, čo umožnilo jemnú kontrolu nad konverziou dokumentov najmä na generovanie konzistentných výstupov z textov zakódovaných v TEI do formátov ako HTML, Markdown alebo LaTeX. Táto skriptovacia vrstva umožnila tímu prispôsobiť transformáciu zložitých štruktúr XML do použiteľných formátov na vedeckú prezentáciu aj výpočtovú analýzu. Táto časť skúma, ako Python a Lua prispeli k modulárnemu, reprodukovateľnému projektu. pracovných postupov, čím sa posilňuje hodnota ľahkého, účelovo vytvoreného skriptovania v digitálnej humanitnej infraštruktúry.

V projektoch digitálnych humanitných vied nie je programovanie ani tak o vytváraní komplexného softvéru ale skôr o navrhovaní flexibilných nástrojov, ktoré pomáhajú skúmať, transformovať a interpretovať údaje. V tomto kontexte sa ukázali obzvlášť užitočné dva jazyky: Python a Lua - každý s vlastnými silnými stránkami a úlohami v rámci pracovného postupu DH.

Python je jedným z najpoužívanejších jazykov v digitálnej humanistike vďaka jeho čitateľnosti, rozsiahlym knižniciam a aktívnej komunite. Je obzvlášť vhodný na úlohy, ako je čistenie textových údajov, analýza frekvencií slov, konverzia formátov súborov alebo vyhľadávanie metadát. Napríklad pomocou knižníc, ako sú lxml alebo BeautifulSoup, môžno efektívne získavať informácie z XML alebo HTML dokumentov , zatiaľ čo nástroje ako Pandas umožňujú výkonnú manipuláciu s údajmi a štatistické súhrny len s niekoľkými riadkami kódu. Python je ideálny na vytváranie opakovateľných, modulárnych skriptov, ktoré sa dajú zdieľať a opätovne používať v ďalších projektoch .

Jazyk Lua je ľahký skriptovací jazyk, ktorý je často súčasťou iných nástrojov . V kontexte DH zažiari, keď sa používa na prispôsobenie pracovných postupov v rámci softvéru, ako je Pandoc - univerzálny konvertor dokumentov, ktorý hrá kľúčovú úlohu v mnohých postupoch transformácie textu. Pomocou jazyka Lua môžno vytvárať kompaktné filtre, ktoré upravujú spôsob konverzie dokumentov, napríklad preformátovanie názvov kapitol, odstránenie poznámok pod čiarou alebo extrakciu špecifických prvkov TEI pred exportom do HTML alebo PDF. Vďaka jednoduchosti jazyka Lua sa ho možno ľahko naučiť pre špecifické, cielené úlohy, najmä pri práci v rámci štruktúrovaných publikačných systémov.

Python a Lua spoločne ponúkajú výkonnú sadu nástrojov: Python na spracovanie údajov a analýzu, Lua na transformáciu a prispôsobenie dokumentov. Zvládnutie dokonca aj základných skriptov v týchto jazykoch môže výrazne rozšíriť možnosti výskumu v oblasti digitálnych humanitných vied a preklenúť tak priepasť medzi tradičnými vedeckými a počítačovými metódami.

## Digitalizácia: Od tlačenej stránky k strojovo čitateľnému textu

The foundation of the corpus-building process began with the digitization of physical novels, many of which existed only in aging print editions or archival microfilm. This phase involved careful selection of source materials based on availability, historical significance, and condition, followed by high-resolution scanning and optical character recognition (OCR). While OCR technologies offer substantial time savings, the process also revealed the limitations of automated text capture when applied to older Slovak orthographies, non-standard typography, or damaged pages. As a result, post-OCR correction—both automated and manual—became a key component of the digitization workflow. This section outlines the practical and methodological considerations that shaped the transition from analog texts to machine-readable data, including the tools, standards, and quality control measures employed to ensure that the digital texts would be suitable for subsequent encoding and analysis.

## Kódovanie: Štruktúrovanie textu podľa schémy TEI

Once the novels were digitized and cleaned, the next step involved enriching the plain text with semantic and structural markup using the Text Encoding Initiative (TEI) guidelines. This phase was central to transforming the corpus into a scholarly resource that could support both humanistic inquiry and computational analysis. TEI encoding allowed for detailed representation of textual features such as chapter divisions, narrative perspective shifts, named entities, quotations, and paratextual elements (e.g., prefaces, footnotes). It also facilitated the inclusion of bibliographic metadata, authorial information, and historical publication context. Balancing descriptive accuracy with encoding efficiency required the development of project-specific schemas and tagging conventions, as well as the use of both automated tagging scripts and manual interventions. This section delves into the rationale behind the encoding strategy, the challenges of modeling 19th- and early 20th-century Slovak prose, and the tools and workflows adopted to ensure consistency and interpretive flexibility.

## Prezentácia: Publikovanie pomocou TEI Publisher

With the corpus fully encoded, the final phase focused on making the material available through a web-based interface that preserved the richness of the TEI markup while offering a smooth, intuitive user experience. For this, TEI Publisher served as the central platform, chosen for its native support of TEI-XML and its flexibility in presenting complex textual structures. The platform enabled not only the display of texts but also faceted browsing, full-text search, and customizable views tailored to different user groups—whether scholars, educators, or general readers. TEI Publisher’s reliance on standards-based technologies like XSLT and REST APIs also allowed for future integration with visualization tools or external datasets. This section discusses the implementation of TEI Publisher in the context of the corpus, detailing how its configuration and extensions were used to bridge the gap between encoded data and accessible digital editions.

Figures

# Bibliografia

Brown, Gordon, ed. *The Universal Declaration of Human Rights in the 21st Century, a Living Document in a Changing World*. Cambridge, [New York]: Open Book Publishers ; NYU Global Institute for Advanced Study, 2016.

“Extensible Markup Language (XML) 1.0 (Fifth Edition).” Accessed April 15, 2025. <https://www.w3.org/TR/REC-xml/>.

Fennell, Philip. “Extremes of XML.” In *XML London 2013 Conference Proceedings*, 80–86. XML London, 2013. <https://doi.org/10.14337/XMLLondon13.Fennell01>.

“History of Unicode.” Accessed April 26, 2025. <https://www.unicode.org/history/>.

“Unicode Standard.” Accessed April 15, 2025. <https://www.unicode.org/standard/standard.html>.

“What Is XML (Extensible Markup Language)?” *WhatIs*. Accessed April 16, 2025. <https://www.techtarget.com/whatis/definition/XML-Extensible-Markup-Language>.

“XSL Transformations (XSLT) Version 2.0 (Second Edition).” Accessed April 22, 2025. <https://www.w3.org/TR/xslt20/>.

1. Text Encoding Initiative [↑](#footnote-ref-30)
2. Predchádzajúce a nasledovné argumenty však poskytujú dôvody v neprospecb týchto formátov aj pre tieto použitia. [↑](#footnote-ref-32)
3. [“Extensible Markup Language (XML) 1.0 (Fifth Edition)”](#ref-extensible2025). [↑](#footnote-ref-33)
4. Ide o univerzálne kódovanie znakov určené na podporu celosvetovej výmeny, spracovania a zobrazovania písaných textov rôznych jazykov a technických disciplín moderného sveta. ([“Unicode Standard”](#ref-unicode2025)) [↑](#footnote-ref-34)
5. [Fennell, “Extremes of XML”](#ref-fennell_extremes_2013). [↑](#footnote-ref-35)
6. [“What Is XML (Extensible Markup Language)?”](#ref-whatisxml2025) [↑](#footnote-ref-36)
7. Technickým detailom implementácie XML sa venujeme nižšie. [↑](#footnote-ref-37)
8. V tomto kontexte by mohlo ísť o význam “tvorca textu, ktorého je označený reťazec časťou” [↑](#footnote-ref-38)
9. XSLT (Extensible Stylesheet Language Transformations) je jazyk pôvodne navrhnutý na transformáciu dokumentov XML do iných XML dokumentov alebo iných formátov, ako je HTML, obyčajný text alebo formátovacie objekty XSL. Tieto formáty možno následne konvertovať do formátov, ako sú PDF, PostScript a PNG. Podpora transformácie JSON a obyčajného textu bola pridaná v neskorších aktualizáciách špecifikácie XSLT 1.0. ([“XSL Transformations (XSLT) Version 2.0 (Second Edition)”](#ref-xslt2025)) [↑](#footnote-ref-40)
10. Postupnosť automatizovaných krokov alebo procesov, ktoré konvertujú údaje z jedného formátu alebo štruktúry do iného. [↑](#footnote-ref-41)
11. Máme tu na mysli to, čo sa v anglickom jazyku označuje ako “plain text”, teda dáta, ktoré obsahujú len reprezentácie znakov čitateľného materiálu bez ich grafickej reprezentácie alebo ďalších objektov (čísiel, s pohyblivovu desatinnou čiarkou, obrázkov, atď.) Niekedy sa síce XML považuje za tzv. bohatý text (“rich text”), keďže okrem reprezentácií znakov čitateľného materiálu obsahuje aj informácie o štruktúre dokumentu alebo informácie slúžiace pre potreby vizuálnej prezentácie textu, ako napríklad, že určitá časť textu má byť v kurzíve ale v určitej farbe, ale podstatné je, že aj tieto informácie majú formu reprezentácií pre človek a počítače čitateľných znakov. [↑](#footnote-ref-44)
12. XML schéme TEI sa venujeme nižšie. [↑](#footnote-ref-46)
13. V tomto prípade ide o kódovanie znakov UTF-8 (*Unicode Transformation Format – 8-bit*) definované už spomínaným štandardom Unicode. V skratke ide o to, že Unicode priraďuje jednotlivým znakom čísla v hexadecimálnej sústave (napríklad U+0041 pre veľké latinizované písmeno A) a UTF-8 priraďuje týmto kódom čísla binárnej sústave. Pre vysvetlenie motivácie tohto dvojitého kódovania pozri ([“History of Unicode”](#ref-unicodehist2025)). UTF-8 je najrozšírenejším kódovaním, keďže podporuje takmer všetky jazyka sveta. [↑](#footnote-ref-48)
14. Verzia 1.0, definovaná v roku 1998, je nejrozšírenejšou a odporúčanou verziou XML. Okrem nej existuje aj verzia 1.1, ktorá sa od predchádzajúcej verzie líši v niekoľkých ohľadoch. Tie tu však nebudeme uvádzať, keďže novšia verzia je málo rozšírená a jej špecifiká pre nás nie sú pre nás podstatné. Ďalšie verzie XML zatiaľ neexistujú. [↑](#footnote-ref-49)
15. Ide o informáciu, či je dokument závislý výhradne od informácií, ktoré sa v ňom náchádzajú (‘yes’) alebo nie (‘no’). [↑](#footnote-ref-50)
16. Niekedy sa na označenie týchto prvkov používa aj výraz “tag”, respektíve “značka”. V ďalšom texte budeme tieto varianty používať zameniteľne. [↑](#footnote-ref-51)