AUTOMATIZOVANÁ KONVERZE A OBOHACENÍ DAT Z DIGITÁLNÍCH KNIHOVEN DO STANDARDU TEI

Boris Lehečka; lehecka@mzk.cz; (Moravská zemská knihovna v Brně)

Účel – Článek představuje výsledky spolupráce českých odborníků v oblasti knihovnictví a digitálních humanitních věd, který zefektivní další využití publikací z digitálních knihoven při primárním a sekundárním výzkumu nebo při trénování kvalitnějších velkých jazykových modelů: automatizovaný proces konverze publikací z digitálních knihoven do standardu TEI včetně obohacení jazykových dat.

Design/Přístup/Metody – Pomocí automatizovaného převodu metadat a obohacování textových dat z digitálních knihoven Kramerius je možné zpřístupnit bohaté kulturní a duchovní dědictví ve standardu TEI, který vyvíjí mezinárodní konsorcium programátorů, knihovních pracovníků a humanitně zaměřených vědců a pedagogů. Autor popisuje data a metadata dostupná díky Standardu Národní digitální knihovny pro jednotlivé publikace, možnosti jejich obohacení pomocí služeb pro zpracování přirozeného jazyka (UDPipe a NameTag) a jejich převod na odpovídající elementy ve standardu TEI. Zaměřuje se na problémy spojené zejména s formátem ALTO a na jejich řešení.

Výsledky – Ověřené postupy a pravidla pro generování validních dokumentů TEI z publikací digitálních knihoven, které byly implementované ve dvou samostatných, volně dostupných aplikacích: DL4DH TEI Converter (využívající systém Kramerius+) a Libri augmentati.

Originalita/Hodnota – Článek popisuje procesy používané při převodu metadat z formátu MODS a ALTO do standardu TEI, které se mohou aplikovat na všechny publikace z digitálních knihoven Kramerius. Díky rozpoznaným entitám v textech publikací je lze snadněji dohledat i z webové aplikace.

Limity – Kvalita výstupu (zejména lingvistická analýza textu) je závislá na vstupních datech (především kvalitě procesu OCR). Některé nedostatky se díky předzpracování vstupních údajů v procesu konverze daří eliminovat.

Klíčová slova – TEI; ALTO; počítačové zpracování přirozeného jazyka; digitální humanitní vědy.

https://doi.org/10.52036/1335793X.2025.2.85-104

ÚVOD

Cílem knihoven je uchovávat a zprostředkovávat kulturní a duchovní dědictví generací minulých i současných pro generace současné i budoucí. Účelem digitálních knihoven je využívat k tomuto cíli počítačové hardwarové i softwarové prostředky, které se opírají o dohodnuté standardy. Cílem badatelů z oblasti digitálních humanitních a společenských věd je využívat při své práci moderní počítačové hardwarové i softwarové prostředky, a to při formulaci výzkumných otázek, shromažďování i analýze badatelského materiálu a nakonec při prezentaci dosažených výstupů ve formě nových poznatků, metadat a obohacených původních údajů.

Následující text vychází z výsledků dosažených v projektu „DL4DH – vývoj nástrojů pro efektivnější využití a vytěžování dat z digitálních knihoven k posílení výzkumu digital humanities“. Z perspektivy počítačového lingvisty popisuje principy i detaily automatizovaného převodu bibliografických dat a zejména obohacení textů o lingvistické údaje a jejich konverze do standardu TEI, za něž byl autor článku v rámci řešeného grantu primárně zodpovědný: bylo potřeba najít pro existující nebo generovaná (meta)data jejich ekvivalent v TEI. Na tyto poznatky a zkušenosti jsem mohl navázat při vývoji desktopové aplikace Libri augmentati[1], která v menším měřítku slouží individuálním badatelům k podobnému účelu jako serverová aplikace DL4DH TEI Converter[2].

Volba standardu TEI jako jednoho z výstupů pro publikace z digitálních knihoven Kramerius byla záměrná: TEI má prostředky pro zachycení metadat, pro zachycení textu, jeho struktury i sémantiky, nabízí prostředky pro odlišení originální podoby předlohy a zásahů editora[3], jedná se o léty prověřený formát spravovaný komunitou, která se podílí na jeho vývoji, propagaci i aplikaci v oblasti výzkumu a výuky. V Česku není moc používaný ani vyučovaný, což může vést k otázce, co má být dřív: zda vejce (poptávka po publikacích ve standardu TEI), nebo slepice (publikace ve standardu TEI)? Pokud budeme podporovat obojí, tj. dostupnost textů a publikací, které dodržují standard TEI, a informovanost badatelů o tomto standardu, dostupných nástrojích včetně metod analýzy nebo prezentace, nebude třeba na tuto otázku v blízké (možná však až vzdálenější) budoucnosti odpovídat. Věřím, že představené metody a výstupy, i když nejsou ideální a bezchybné, přitáhnou pozornost profesních a badatelských kruhů k tomuto standardu.

1 VÝBĚR Z LITERATURY A APLIKACÍ

Obecný základ převodu dat z knihovnických systémů do standardu TEI představuje dokument Best Practices for TEI in Libraries (TEI Consortium, September 2018), který obsahuje doporučení pro konverzi z formátu MARC na prvky standardu TEI a dále popisuje několik úrovní podrobnosti značkování samotného textu publikace: od čistě automatického OCR (Level 1) přes minimální strukturování s navigačními elementy (Level 2) až po plně samostatný elektronický text s logickou strukturou a základní typografickou analýzou (Level 3). Úroveň 4 přidává podrobnou popisnou informaci o obsahu (členění na titulní stranu, oddíly, odkazy, případně scény u dramatických textů apod.) a v případě úrovně 5 i pokročilé jazykové, prozodické a textověkritické anotace, které ale nejde automatizovat, neboť vyžadují aktivní zásahy badatele.

Anonymní dotazníkový průzkum z roku 2013 (Dalmau a Hawkins 2014) na téma využití standardu TEI v knihovnách, do kterého se zapojilo minimálně 41 institucí (zbylých 57 respondentů se nepodařilo identifikovat), ukázal, že se převod do standardu TEI využívá zejména u specializovaných sbírek a jedinečného obsahu, nikoli pro obecně dostupné publikace.

O zachycení výstupu OCR (původně ve formátu ALTO nebo PAGE XML) pomocí standardu TEI se pokusil Scheithauer et al. (2024). Burnard (2021) navrhuje používat TEI jako výchozí formát pro sdílení a výchozí bod pro další transformace, konverze a obohacení, které mohou generovat další formáty.

Obecně využívají projekty zaměřené na převod digitalizovaných publikací do standardu TEI digitální knihovny hlavně jako zdroj bibliografických a obrazových dat. Pro kvalitnější rozpoznání textu se používají specializované platformy, např. eScriptorium[4] nebo Transcribus[5] (Scheithauer et al. 2024). Pro generování dokumentu TEI slouží transformace vygenerovaných výstupů ve formátu ALTO nebo PAGE XML s využitím transformací XSLT[6] nebo skriptů v programovacím jazyce Python[7] (Chiffoleau 2024). S dokumenty TEI se dále pracuje prostřednictvím dalších specializovaných nástrojů, jako je Leaf-WRITER[8] (on-line editor, ruční anotace pojmenovaných entit), TEITOK[9] (automatizovaná tokenizace[10], lemmatizace, morfologická anotace, prezentace) nebo TEI Publisher[11] (ruční a poloautomatická anotace pojmenovaných entit, prezentace).

Budování lemmatizovaných a lingvisticky anotovaných textů z digitálních knihoven připomíná budování jazykových korpusů. Rozdíl spočívá v jejich zpracování i zaměření: jazykový korpus představuje „rozsáhlý soubor autentických textů (psaných nebo mluvených) převedený do elektronické podoby v jednotném formátu tak, aby v něm bylo možné jednoduše vyhledávat jazykové jevy“[12]. Na rozdíl od digitálních knihoven korpusy často nemají přímou vazbu na digitalizát, z něhož vycházejí: v některých případech jde o tzv. born-digital texty bez klasické tištěné předlohy (nebo tato vazba není zaznamenána, např. u elektronické verze periodik). Jazykové korpusy bývají co do velikosti omezené a vyvážené, tj. obsahují pouze výběr ze všech dostupných textů, ale jednotlivé kategorie textů (např. beletrie, próza, publicistika, drama), které se stanovují na základě plánovaného využití korpusu, bývají zastoupeny tak, aby jejich poměrné zastoupení v korpusu odpovídalo jejich zastoupení v celku. V případě autorsky chráněných děl bývají ve veřejně dostupných korpusech náhodně prohozené odstavce, případně věty. Knihovny usilují o zveřejnění všech dostupných publikací, včetně např. jednotlivých vydání téhož díla, a to se zachováním co nejvěrnější podoby originálu. Na druhou stranu určité fáze zpracování textů, které jsou tématem tohoto článku, například lingvistická anotace a použité nástroje, zůstávají stejné, i když se v detailech můžou lišit (např. použitý repertoár morfologických značek nebo způsob jejich zachycení pomocí atributů či vertikály[13]).

Existují však i korpusy tvořené dokumenty podle standardu TEI, například *Deutsches Textarchiv* (DTA) (Berlin-Brandenburgische Akademie der Wissenschaften 2007–2025), textový archiv německých, převážně tištěných textů z období 17.–19. století, EarlyPrint Library

(EarlyPrint Library [bez data]) věnující se přepisům zejména anglickojazyčných starých tisků z let 1470 až 1800[14], uložených v téměř 150 knihovnách a digitalizovaných v rámci projektu *Text Creation Partnership* (TCP) (Text Creation Partnership [bez data]), nebo francouzské úložiště světové vědecké literatury ISTEX (ISTEX [bez data])[15]. Tento přístup rovněž ve větší míře využívají projekty zaměřené na zpracování a publikaci užšího okruhu děl, obvykle orientovaných na určité teritorium, jazyk, časové období nebo konkrétního autora.[16]

Automatizované konverzi a obohacení dat z digitálních knihoven se věnuje Lhoták (2020, s. 29) v rámci představení úvodní fáze projektu DL4DH: snaží se o využití existujících dat (ALTO, MODS aj.), obohacení textu pomocí externích služeb pro lemmatizaci, morfosyntaktickou anotaci a identifikaci pojmenovaných entit a následné sloučení těchto dat do dokumentu TEI. S drobnými úpravami v důsledku konkrétní implementace popsaných principů (např. generování elementu **<teiHeader>** místo původního **<text>**) přichází Lehečka et al. (2022, s. 33). Práce podrobněji popisuje využívané externí služby (Lehečka et al. 2022, s. 27–28) a uvádí ukázku ve standardu TEI; podrobnější informace o volbě konkrétních elementů TEI chybí (s výjimkou odkazu na tabulku s převodníkem elementů XML ze služby NameTag).

Vybrané detaily procesu konverze (identifikace odstavců v textových blocích ALTO) a obohacení (použití výstupu služby UDPipe jako vstupu pro službu NameTag) zachycuje prezentace na toto téma na konferenci TEI 2025 v Krakově (Lehečka 2025a).

2 VÝCHODISKA AUTOMATIZOVANÉ KONVERZE A OBOHACOVÁNÍ

Cílem projektu DL4DH byl „vývoj nových funkcí a nástrojů, které umožní extenzivní využití a vytěžování dat z digitálních knihoven pro potřeby digitálního humanitního výzkumu, a současně příprava aplikovaných vědeckých výstupů využívající tyto nové možnosti“ (Lhoták 2020, s. 26). Při vývoji se využila digitální knihovna na bázi sytému Kramerius[17], který operuje pouze nad jednou sbírkou publikací (obvykle) z jedné instituce. V tomto systému se pro data využívají různé formáty na bázi značkovacího jazyka XML, každý za specifickým účelem: FOXML[18] (Fedora Object XML; popisná, administrativní a strukturální metadata), MODS[19], Dublin Core[20] (popisná metadata), JSON-LD (struktura a rozložení digitalizovaného dokumentu dle specifikace IIIF Presentation API v3) a ALTO (rozložení textu na stránce). Vedle toho ještě obrazová data ve formátu JPEG a rozpoznaný text (odvozený z formátu ALTO) v podobě prostého textu. Zveřejňují se pouze takové publikace, jejichž data a metadata vyhovují Standardům Národní digitální knihovny (NDK) (Národní knihovna ČR 2024), přičemž pro různé typy publikací se vytvářejí tzv. Definice metadatových formátů (DMF), z nichž se v roce 2024 osamostatnil popis pro rozpoznaný text (ALTO XML a TXT OCR) (Hutař et al. 2024).

2.1 ALTO (ANALYZED LAYOUT AND TEXT OBJECT)

Dokumenty standardu ALTO (Library of Congress [bez data])[21] nesou informaci o rozmístění textu a netextových objektů na stránce a výsledky rozpoznání tištěného, popř. strojopisného nebo rukopisného textu, tj. OCR (Optical character recognition 2001-), popř. HTR (Handwriting recognition 2001-). Přestože existují poměrné jednoduché XSLT transformace nebo skripty, které vygenerují prvky dokumentu TEI pro text i obraz, bude vhodné se tomuto standardu věnovat podrobněji kvůli úskalím, s nimiž je při transformaci potřeba počítat, a vzhledem k možnosti označovat pojmenované entity i další součásti textu, které se postupně budou prosazovat, např. díky podpoře v poslední Definici metadatových formátů pro ALTO XML (Hutař et al. 2024).

Dokument ALTO XML zachycující digitální předlohu je rozdělen na několik základních částí: nastavení (měrné jednotky, metadata); styly používané strukturními prvky na stránce (např. písmo, zarovnání odstavce); značky definující vlastnosti, na něž je možné odkázat (např. pojmenované entity, role); pořadí čtení segmentů na stránce, pokud není kontinuální; samotná stránka tvořená okraji a potištěnou plochou, na níž se vyskytují jednotlivé rozpoznané části (bloky textu, ilustrace, grafické elementy apod.). Ne všechny tyto prvky musí být přítomné a ne všechny zatím DMF podporuje.

V úvodním popisu dokumentu (**<alto>/<Tags>**) lze použít prvky, které nesou podrobnější informaci o roli nebo významu rozpoznaných částí na stránce. Tímto způsobem je možné identifikovat např. razítka, reklamy (**<LayoutTag>**), nadpisy, živá záhlaví (**<StructureTag>**), autora, vydavatele (**<RoleTag>**), jména osob, názvy institucí (**<NamedEntityTag>**), popř. další zaznamenání hodný prvek (**<OtherTag>**).[22]

Mezi hlavní prvky ALTO pro zachycení rozpoznaného textu a dalších objektů patří následující:
Element **<Styles>** obsahuje prvky pro definici stylů používaných na stránce, a to zvlášť na úrovni odstavců (**<ParagraphStyle>**) a na úrovni úseků textu (**<TextStyle>**), na tyto styly se z objektů na stránce odkazuje pomocí identifikátoru. Podřazené objekty v tomto případě dědí vlastnosti z objektů nadřízených.[23]

Na úrovni formátování odstavců lze zachytit následující vlastnosti: zarovnání (@ALIGN, tj. vlevo, vpravo, na střed, do bloku), odsazení zleva (@LEFT), popř. zprava (@RIGHT), řádkování (@LINESPACE) a odsazení prvního řádku (@FIRSTLINE).

Pro podchycení formátování textových úseků jsou k dispozici tyto vlastnosti: název fontu (@FONTFAMILY, např. Times New Roman), typ (@FONTTYPE, tj. patkový, bezpatkový), šířka (@FONTWIDTH, tj. pevná nebo proporcionální), velikost (@FONTSIZE; v bodech, tj. 1/72 palce), barva (@FONTCOLOR) a styl fontu (@FONTSTYLE, např. tučné nebo kurzivní písmo, horní index apod.).

Výchozí prvek, který nese textovou informaci, představuje element **<TextBlock>**. Definuje sdílené typografické vlastnosti na úrovni bloku a je tvořen jedním nebo několika samostatnými řádky, elementy **<TextLine>**. Podřazenými prvky **<TextLine>** jsou elementy **<String>** (text tvořený písmeny a interpunkčními znaménky) a prázdné prvky **<SP>** (text tvořený prázdným prostorem, tj. mezery a tabulátory), případně **<HYP>** (obsahuje spojovník, popř. jiné dělicí znaménko na konci řádku).

Element **<String>** obsahuje rozpoznaný text v atributu @CONTENT. Vedle toho je k dispozici také atribut @SUBS_CONTENT, jehož obsah tvoří upravený rozpoznaný text, např. text s rozepsanou zkratkou nebo kompletním slovem, které je v předloze rozděleno spojovníkem. Pomocí hodnoty atributu @SUBS_TYPE lze specifikovat, o jaký typ úpravy se jedná. Pomocí atributu @WC lze zachytit míru pravděpodobnosti (hodnota v rozmezí 0–1), s níž rozpoznávací software text danému úseku přiřadil. Pravděpodobnost správného přiřazení jednotlivých liter se zachycuje v atributu @CC: pro každý znak se uvede číselné hodnocení na škále od 0 (jisté) do 9 (nejméně pravděpodobné), čísla se oddělují mezerou. Atribut @CS slouží k zachycení faktu, zda došlo k ruční opravě rozpoznaného textu.

V případě potřeby je možné v elementu **<String>** zachytit také alternativní rozpoznané podoby, a to pomocí vnořeného elementu **<ALTERNATIVE>** a jeho atributu @PURPOSE, tentokrát však již bez určení míry pravděpodobnosti. Pomocí vnořených elementů **<Glyph>** lze rozpoznaný text dále rozdělit na jednotlivé glyfy (písmena, spřežky) a určit míru pravděpodobnosti jejich přiřazení pomocí atributu @GC. Také u této nejmenší textové jednotky lze zachytit alternativní podobu, a to ve vnořeném elementu **<Variant>** (např. pro případy typu „m“ a „rn“).[24]

Jazyky

Pro správné rozpoznání textu pomocí OCR, ale i jeho následnou analýzu pomocí nástrojů pro zpracování přirozeného jazyka (NLP) je důležitá informace o jazyce, v němž jsou dokument nebo jeho dílčí části vytvořeny.

Standard ALTO zavedl označení jazyka na úrovni textového bloku (**<TextBlock>**), řádku (**<TextLine>**) a slova (**<String>**) pomocí atributu @LANG, a to od verze 2.1, od verze 4.4 jej lze použít i pro celou stranu (**<Page>**). Do verze 2.0 se používal atribute @language. Na hodnotu atributu se kladou stejná omezení jako na obecný atribut @xml:lang, která vycházejí z doporučení BCP 47.[25] Analýza materiálu z digitálních knihoven Kramerius dokládá použití atributu @language pouze u elementu **<TextBlock>**, u nižších jednotek se jazyk nezachycuje.

Formátování textu

Jelikož mají v digitálních knihovnách Kramerius identifikátory stylu (tj. hodnota atributu @ID u elementů **<ParagraphStyle>** a **<TextStyle>**) podobu jedinečných identifikátorů (GUID) bez vazby na samotné kvality formátování, může mít tentýž formát při zpracování jednoho dokumentu odlišné identifikátory na různých stranách, takže je při slučování stran potřeba vždy vycházet z hodnot konkrétního stylu.

K dalším komplikacím dochází, když software pro rozpoznání textu identifikuje v originále velké množství formátování (např. několik velikostí písma s odstupňováním po 0,5 bodu), třebaže se při sazbě použil jenom omezený repertoár písem. Tato rozkolísanost je daná kvalitou originálu i digitální kopie, ale může mít dopad na další automatizované zpracování rozpoznaného textu, které počítá s určitými kvalitami typografického zpracování, jako je tomu např. u slovníků.

Umístění a tok textu

U složitějších publikací, jako jsou ilustrované knihy nebo noviny a časopisy, je důležité zachytit zamýšlený tok textu, který bývá přerušen ilustracemi, inzeráty nebo bloky samostatného, nesouvisejícího textu, v některých případech může text navazovat i o několik stránek dále. Z hlediska dalšího automatizovaného zpracování a jazykové analýzy je důležité zachovat myšlenkovou i syntaktickou návaznost rozdělených částí textu. I když DMF následující elementy standardu ALTO bohužel nepovoluje, uvádíme je pro úplnost. Od verze ALTO 4.3 lze v metadatech ke stránce (**<alto>/<ReadingOrder>**) definovat pomocí odkazů na textové bloky jejich správné kontinuální pořadí (**<OrderedGroup>**), stejným

způsobem lze zachytit i bloky textu, které spolu nesouvisí (**<UnorderedGroup>**).

Dělení slov

V současném psaném jazyce se pro tyto účely v latince používá spojovník (tiret, divis), který mohl mít v dřívějších dobách i různou podobu (připomínající např. rovnítko[26]). Nerozpoznaný spojovník má dopad na věcnou i gramatickou správnost delšího úseku.

Standard ALTO používá pro zachycení spojovníku samostatný element **<HYP>**, zajímavé přitom je, že má povinný atribut `@CONTENT`, který na základě analýzy dat z digitálních knihoven Kramerius obvykle obsahuje hodnotu „175“, což je desítková hodnota znaku pro logickou negaci (¬), asi v 10 % případů obsahoval atribut přímo znak pro tzv. volitelnou pomlčku (U+00AD) a pouze v těchto případech se u předchozího a následujícího slova (**<String>**) objevil atribut `@SUBS_CONTENT`, který obsahuje kompletní podobu rozděleného slova.

Jazyková analýza textových dat

Obsah digitalizovaných děl, tj. jejich text, je při zpracování možné opatřit lemmatizací, morfologickou anotací a identifikací pojmenovaných entit. Jedná se o výsledky analýz pomocí nástrojů počítačového zpracování přirozeného jazyka (NLP), které mohou sloužit jako vhodný základ pro jakékoliv další zpracování materiálu pro konečnou analýzu. Pro tento účel byly vybrány dvě aplikace, které vyvíjí a provozuje formou webové služby velká výzkumná infrastruktura LINDAT/CLARIAH-CZ a které kvalitně zpracovávají česky psaný text.

UDPipe (Straka a Straková 2016) vstupní texty tokenizuje, lemmatizuje a opatřuje morfologickou a syntaktickou anotací. NameTag (Straka a Straková 2014) dokáže v textu rozpoznat a označit pojmenované entity. Oba lingvistické nástroje podporují několik jazyků a několik formátů pro vstupní i výstupní text. Zpracování textu probíhá na základě zvolených natrénovaných jazykových modelů.

3 METODOLOGIE

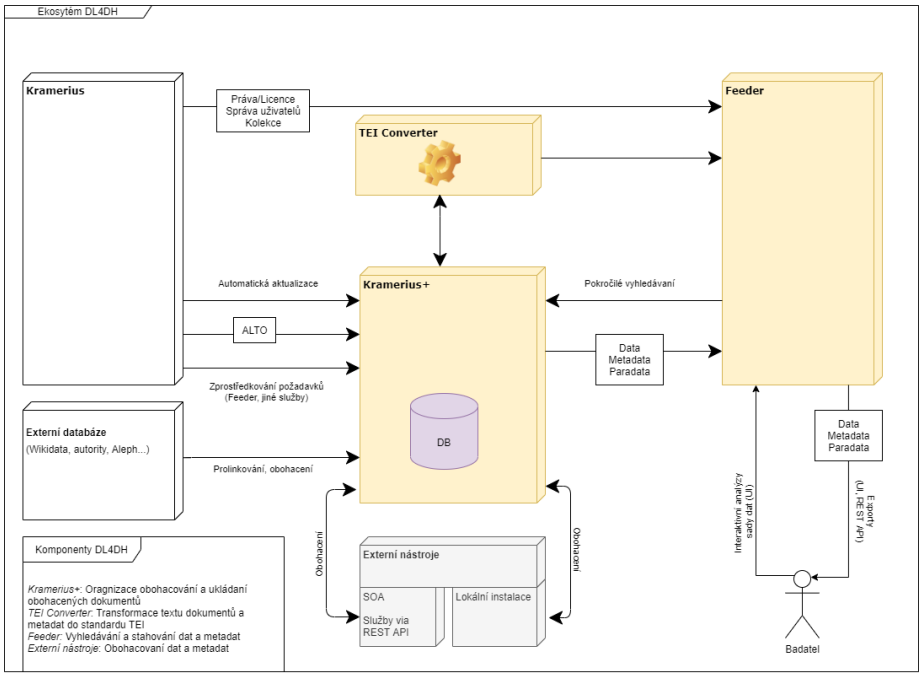
Sekundárním cílem projektu DL4DH bylo posílit výzkum v oblasti digitálních humanitních věd, proto jsme volili prostředky, na které jsou badatelé z tohoto okruhu zvyklí: data ve strukturované a standardizované podobě, přístupná přes uživatelské i programové rozhraní. Výzkumníci tak můžou pracovat s nástroji, které jsou pro oblast jejich bádání typické a ideální, aniž by se museli zabývat nadbytečnou konverzí dat.

Naší snahou bylo dodat poskytovaným textům i další přidanou hodnotu. Rozhodli jsme se pro lemmatizaci, morfosyntaktickou anotaci a identifikaci pojmenovaných entit, protože se jedná o minimální společný základ pro velkou část výzkumů, který zpřesní prohledávání a identifikaci relevantních pasáží. A pokud tato data nejsou pro další výzkum potřebná, lze je snadno odstranit, případně negenerovat. Ostatní zvažovaná obohacení (např. identifikace bibliografických citací) buď přesahovala časové a finanční možnosti projektu, nebo se ukázaly jako specifické pro užší segment badatelů.[29]

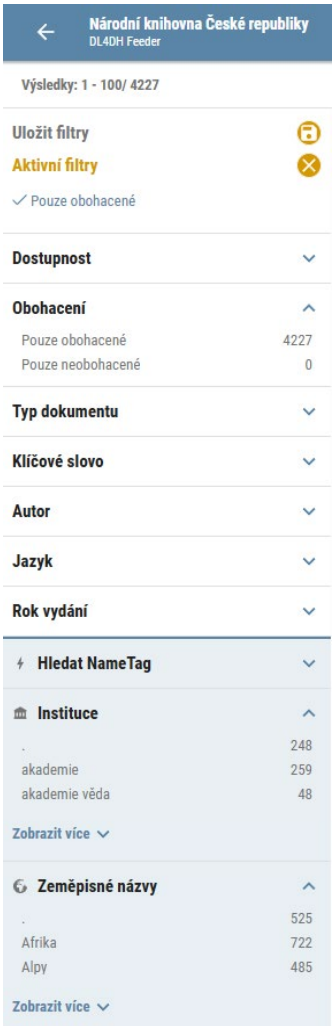
V ideálním případě bude možné zpřístupnit tímto způsobem všechna textová data uložená v digitálních knihovnách Kramerius[30], v praxi narážíme na různé problémy, zejména u samotného textu publikací: nedostupná data (např. formát ALTO), nekvalitně rozpoznaný text, použití staršího pravopisu, který nelze zpracovat momentálně dostupnými nástroji pro zpracování přirozeného jazyka. Z dlouhodobé perspektivy lze tyto překážky postupně odstranit: provést nové rozpoznání textu (OCR) s kvalitnějšími nástroji, aplikovat na text před jazykovou analýzou transkripční pravidla nebo natrénovat adekvátní jazykové modely pro nástroje NLP. Z těchto důvodů jsme se v projektu nevěnovali evaluaci chybovosti vstupních dat (ve formátu ALTO) a jejich dopadu na kvalitu výstupu.

Aby bylo možné bibliografická data, paradata o zpracování i samotné texty vyžívat při výzkumu, který primárně využívá počítačové zpracování, zvolili jsme několik nových standardizovaných formátů, do kterých se uvedená data převádějí: CSV/TSV, JSON a TEI. Přidanou hodnotou, která v digitálních knihovnách Kramerius zatím nebyla dostupná, je obohacení textu publikace o anotaci jazykových jevů, jako jsou slovní druhy, morfosyntaktické kategorie (pád, číslo, vid aj.) nebo rozpoznané entity (osoby, místa apod.).

O generování dokumentů TEI se stará v rámci projektu DL4DH vyvinutá aplikace Kramerius+[31], která na základě metadat ze systému Kramerius odesílá textová data na externí služby a získané výsledky pomocí aplikace DL4DH TEI Converter[32] transformuje na kompletní dokument TEI s obohacenými daty. Celé řešení komunikuje s jednou instancí Krameria a je určeno pro serverové nasazení. Aplikace DL4DH Feeder agreguje data z Krameria a Krameria+ a slouží k vyhledávání, prohlížení a stahování dat prostřednictvím webového prohlížeče. Viz schematické zobrazení komponent a jejich propojení na obrázku 1.



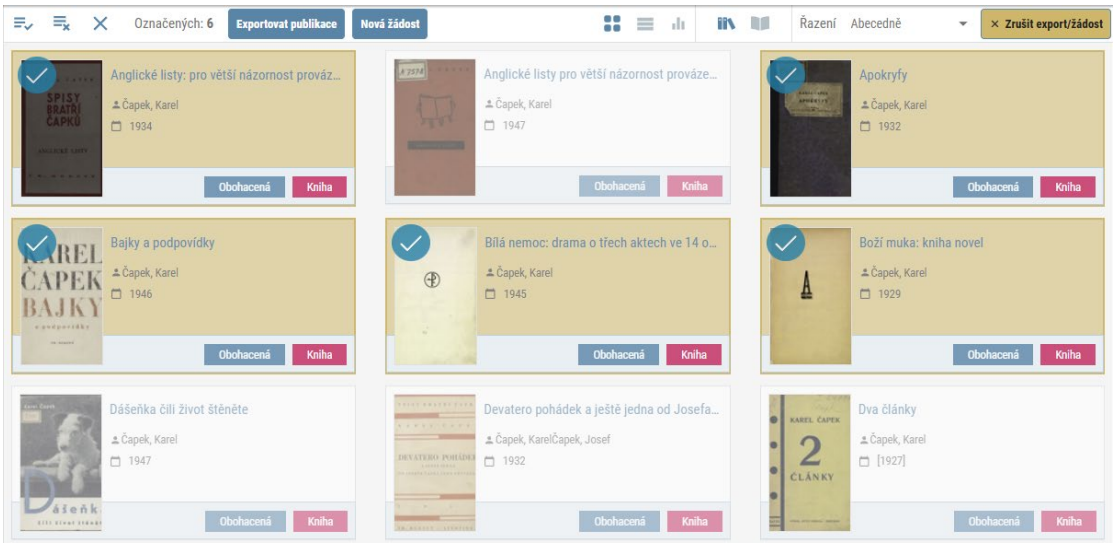
Obrázek 1 Základní prvky řešení DL4DH



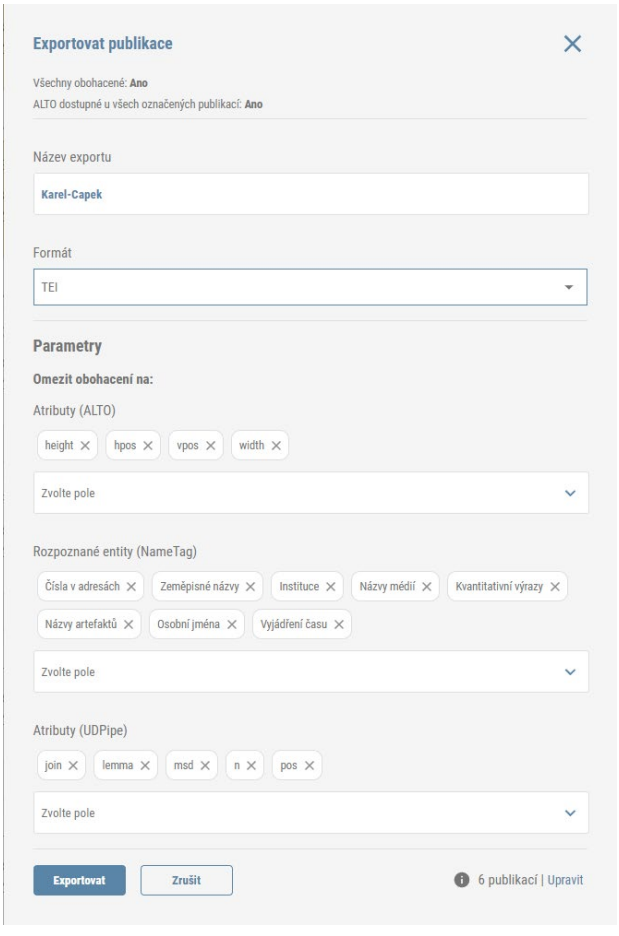
Obrázek 2 Filtry rozpoznaných entit v textech publikací

Pro zobrazení obohacených publikací ve webové aplikaci a jejich stažení (např. ve formátu TEI) jsou potřeba následující kroky: knihovna musí zprovoznit aplikace DL4DH Feeder a Kramerius+, kterou pomocí konfigurace propojí s vlastní instancí systému Kramerius a externími službami pro obohacení. Správce digitálních sbírek vybere publikace a zadá požadavek na jejich zpracování, který obsahuje identifikátory publikací a parametry obohacení. Po úspěšném zpracování správce vybere publikace, exportuje je do formátu TEI a povolí jejich zveřejnění v aplikaci DL4DH Feeder.[33] Jednotlivé kroky ilustruje Lehečka (2025b, s. 14–18).

Zájemce o data využije webovou aplikaci pro vyhledání publikací, jejich výběr (viz obrázek 3) a žádost o export. Vyhledávat lze na základě obvyklých bibliografických údajů, ale i pomocí textového obsahu, jako jsou osoby nebo místa, o kterých se v textu zveřejněných publikací píše (viz obrázek 2). U exportu do formátu TEI lze navolit parametry, které má výsledný dokument obsahovat, tj. např. pouze lemmata bez morfosyntaktické anotace nebo kompletní sada údajů (viz obrázek 4). Požadované dokumenty si uživatel nakonec stáhne zazipované. Viz též Lehečka (2025b, s. 5–12).



Obrázek 3 Výběr publikací pro export



Obrázek 4 Nastavení exportu do standardu TEI

3.1 PŘEVOD DO XML PODLE STANDARDU TEI

Dokument TEI se vytvoří transformací bibliografických údajů z formátu MODS (Library of Congress, 2024) a sloučením tří dokumentů ze samostatných transformací: transformací dokumentu ALTO a transformací výsledků volání externích služeb NameTag a UDPipe.

Hlavička dokumentu TEI obsahuje bibliografická data (**<titleStmt>**, **<title>**, **<author>**, **<extent>**, **<publStmt>** apod.), prvek **<facsimile>** obsahuje elementy pro jednotlivé stránky (**<surface>**) a dílčí rozpoznané oblasti (**<zone>**) v původním obrázku. Text je rozdělen na odstavce (**<p>**) a věty (**<s>**), jednotlivá slova (**<w>**) používají atributy (`@lemma`, `@msd`) pro lemmata a morfologické kategorie. Hlavička konečného dokumentu obsahuje kategorie rozpoznaných entit (v prvku **<textClass>**) a také informace o nástrojích použitých při obohacování: název aplikace, použitý jazykový model a licenční ujednání pro výstup (**<appInfo>** a **<availability>**).

3.2 DÍLČÍ KROKY

Při generování dokumentu TEI je potřeba vycházet z formátu, který popisuje celou publikaci. V případě digitální knihovny Kramerius je to popis využívající schéma FOXML. Díky tomu je možné zjistit bibliografické informace o publikaci, ale i fyzickou strukturu dokumentu (počet stran, jejich typ a označení) a ve spojení s programovým rozhraním (REST API) i uložení souvisejících souborů, zejména obrázků a dokumentů ALTO.

Bibliografické údaje publikace obsahuje element **<datastream** `ID="BIBLIO_MODS"`**>** ve formátu MODS. V ele-

mentech **<datastream ID="RELS-EXT">**, popř. **<datastream ID="RELS-EXT.5">** jsou ve formátu RDF zachyceny identifikátory jednotlivých stran dokumentu, a to v elementu **<hasPage rdf:resource="info:fedora/uuid:73648039…">**.

Následně je pomocí volání odpovídajících koncových bodů REST API nutné zjišťovat, zda je dostupný konkrétní formát pro jednotlivé stránky, tj. např. obrázek, prostý text nebo ALTO. V této fázi vznikne základní kostra dokumentu s metadaty, uspořádáním textu (posloupností jednotlivých stránek) a dostupnými formáty pro dílčí strany.

Pro vytvoření hlavičky dokumentu TEI s bibliografickými daty se díky své detailnosti hodí standard MODS (Metadata Object Description Schema), jehož data tvoří součást elementu **<datastream ID="BIBLIO_MODS">**. Mezi jednotlivými elementy tohoto standardu (**<titleInfo>**, **<languageTerm>**, **<publisher>** aj.) a odpovídajícími prvky v dokumentech TEI (**<title>**, **<language>**, **<publisher>** atp.) existuje relativně jednoduchý převodník (viz tabulka 2), který lze definovat např. v transformační šabloně XSLT.[34]

V dalším procesu je vhodné zaměřit se na zpracování samotného textu a jeho obohacení. Pokud má být obsah dokumentu TEI provázaný s konkrétními oblastmi, kde se vyskytuje text nebo grafika, musí být k dispozici dokument ALTO. Pokud je dostupný pouze prostý text, bude možné přiřadit ke konkrétní stránce její digitální obrázek, ale živá záhlaví, odstavce, řádky nebo slova lokalizovat na stránce možné nebude.

Pro obohacení textu o identifikované entity a morfosyntaktickou analýzu je potřeba získat samotný text, který bude podroben analýze externími službami. Také zde je vhodnější vycházet z formátu ALTO, který by měl obsahovat informace o dělení slov na konci řádku (v takovém případě je vhodné rekonstruovat celé slovo). Data, se kterými jsme pracovali, nebyla ve zpracování jednotná (viz výše), proto je vhodné při přípravě textu pro jazykovou analýzu zařadit kontrolu dělení slov.

Z hlediska textové analýzy se u formátu ALTO jako nevýhodný jeví fakt, že identifikuje bloky textu (**<TextBlock>**) na základě formální shody (odsazení z levé/pravé strany, výška řádku apod.), takže několik odstavců na stránce, které dodržují stejná typografická pravidla, tvoří jeden textový blok o několika řádcích a bez další analýzy není možné dílčí odstavce v tomto bloku identifikovat. Pro tyto případy je možné použít následující postup:[35]

- pro každý textový řádek (**<TextLine>**) v bloku zaznamenat informaci o horizontální pozici a šířce řádku;
- dopočítat další hodnoty: počet slov (**<String>**), počet mezer (**<SP>**), počet znaků (tj. na základě textového obsahu všech podřízených prvků **<String CONTENT="…">**), zda řádek začíná velkým písmenem a zda končí spojovníkem;
- vypočítat průměrné hodnoty na základě všech řádků v textovém bloku a porovnat je s údaji pro jednotlivé řádky;
- na základě výsledků porovnání identifikovat počáteční a koncové řádky odstavce;
- s využitím identifikovaných koncových a počátečních řádků rozdělit textový blok do odstavců.

Text z takto vyčleněných odstavců je možné po jednotlivých odstavcích posílat na další analýzu externími jazykovými službami (aby šlo snadno vrátit data na původní místo na stránce).

Při volání externích služeb pro obohacení textu (v našem případě UDPipe a NameTag) je potřeba dbát na to, aby ke vstupnímu textu přistupovaly stejným způsobem, zejména pokud jde o rozdělení textu na tzv. tokeny, což jsou jednotky, které NLP zpracovává[36]. Při testování služeb NameTag a UDPipe se ukázalo, že obě aplikace pracují s prostým textem na vstupu, ale jejich segmentace na tokeny může mít různé výsledky. Těmto problémům se lze vyhnout v případě, že aplikace umí pracovat se vstupním formátem, který zachovává rozdělení textu na tokeny. Jelikož je rozdělení na tokeny vázáno na konkrétní jazyk, je vhodné tuto segmentaci svěřit specializovanému nástroji. V našem případě nejprve odesíláme prostý text na morfosyntaktickou analýzu (nástroj UDPipe), který zpracuje výstup ve formátu CoNLL-U[37], jenž obsahuje jednotlivé tokeny a jejich charakteristiky na samostatných řádcích (navíc také obsahuje rozčlenění na věty). Naštěstí s tímto formátem umí pracovat služba NameTag na vstupu, takže na svém výstupu vrací text ve formátu XML se stejným počtem tokenů jako služba UDPipe. Díky tomu je pak možné sloučit morfosyntaktické a sémantické informace z obou služeb do jednoho dokumentu.

Služba NameTag používá vlastní repertoár značek a atributů pro označování entit, dat, adres apod., takže bylo nutno vytvořit převodník na odpovídající elementy a atributy ze standardu TEI P5, který

zachycuje tabulka 1. Při volbě korespondujících elementů jsme vycházeli z následujících zjištění a zásad:

- Nástroje pro rozpoznávání entit definují svou vlastní taxonomii jednotlivých druhů rozpoznávaných entit; repertoár se může měnit i u téhož nástroje, např. u NameTagu mezi verzemi 1.0 a 2.0.
- Při práci s TEI je vhodnější využívat prostředky (elementy), které nabízí tento standard, a nesnažit se přizpůsobit TEI takovému pojetí, které má software pro rozpoznávání entit. (Zachování původní klasifikace je však možné – a nejspíš i žádoucí.)
- Standard TEI používá vlastní taxonomii (v podobě repertoáru elementů) pro zachycení entit (např. **<address>**, **<num>**, **<date>**); tento repertoár se postupně specializuje: existují speciální značky pro jména osob, míst, organizací apod.; např. **<objectName>** se objevil teprve v lednu 2019.
- Obsah specifických elementů (např. u **<persName>**) je obvykle identický s obecným elementem **<name>**, tj. repertoár zanořených elementů bývá stejný.
- Výhoda specifičtějších elementů spočívá v tom, že je lze pomocí atributu `@type` blíže specifikovat, zatímco u obecnějšího elementu **<name>** se tento atribut využije pro základní rozlišení pojmenování. Srov. např.:
  <name type="person"><name type="forename" subtype="abbreviated">J.</name></name>
  a
  <persName><forename type="abbreviated">J.</forename> </persName>
- TEI se snaží pro obecnější prvky, které např. NameTag považuje za entity, definovat samostatné elementy, např. **<num>** (číslo), **<date>** (datum), **<street>** (ulice) apod. Proto je potřeba při převodu z taxonomie programů pro rozpoznávání entit do standardu TEI vybírat nejvhodnější elementy: pokud používáme **<num>** pro čísla, je na místě použít např. **<orgName>** pro názvy organizací.
- Použití specifičtějších prvků má výhodu v tom, že jde o větší míru standardizace: zatímco názvy elementů se nemění, obsah atributu `@type` obvykle není pevně stanoven. Lze to sice omezit pomocí vlastní úpravy TEI (s využitím ODD), vždy to ale bude méně standardizované než v případě elementů (můžeme se rozhodnout např. pro hodnoty `@type` typu *pers*, *org*, nebo *person*, *organisation*; uživatel bude muset tuto konvenci znát).

Hlavní princip naší transpozice do standardu tedy spočívá v tom, že pokud v TEI existuje pro entitu specifický element, použije se, teprve pokud neexistuje, rozlišujeme entity pomocí hodnot atributů.

Všechny kroky, které se při generování dokumentu ve formátu TEI dějí, by se měly zaznamenat v samotném dokumentu TEI, a to v části pro automatizované zpracování pomocí aplikací (**<appInfo>** a **<availability>**). Tyto údaje je možné získat jednak z nastavení pro externí služby, jednak z údajů, které vrátí samotné externí služby.

3.3 VÝSLEDEK TRANSFORMACE A OBOHACENÍ

Vyjdeme-li z obecných doporučení pro knihovny (TEI Consortium, September 2018) a porovnáme je s výsledky aktuální implementace obohacení a konverze do TEI, můžeme konstatovat, že výstupy odpovídají kombinaci prvků z úrovní 1 (informace o digitální kopii), 2 (text, textové bloky bez označení sekcí) s dílčími prvky z úrovní 4 (identifikace pojmenovaných entit) a 5 (lemmatizace a morfosyntaktické značkování). Ve výstupech není podrobné rozčlenění na specifické části textu (nadpisy, bibliografie, poznámky pod čarou, nadpisy, tabulky, seznamy, scény, strofy, verše apod.), tj. prvky úrovně 3.

V rámci zkušebního provozu projektu DL4DH se z výběru 7850 volně dostupných monografií podařilo transformovat 7115 publikací, u 141 transformace selhala kompletně a u zbytku (594) pouze částečně (obvykle při generování výstupu podle standardu TEI).

3.4 ÚSKALÍ TRANSFORMACE DO TEI

Při převodu do formátu TEI se objevují i další problémy, které je potřeba vzít do úvahy a které mohou mít různá řešení.

Při zpracování digitalizované publikace můžeme narazit na chybějící data. Jednak můžou být nedostupná pouze dočasně (výpadek externí služby nebo internetového připojení), jednak můžou zcela chybět (text není dostupný ve formátu ALTO, ale pouze jako prostý text), jednak můžou být data poškozená (např. nevalidní dokument XML).

U digitalizovaných dokumentů, na rozdíl od born-digital dokumentů, záleží na kvalitě předlohy a na kvalitách aplikace pro rozpoznávání textu (OCR). Není výjimkou, že výstupy z těchto aplikací obsahují chyby typu špatně rozpoznaných znaků, nejednotného zpracování dělení slov na konci řádku nebo nerozpoznaného záhlaví a zápatí.

Tyto faktory pak mají vliv na sestavení textu do původní podoby a jeho následnou analýzu externími aplikacemi pro zpracování přirozeného jazyka.

U jazykové analýzy je potřeba používat jazykové modely určené pro konkrétní jazyk vstupního textu (jiná bude analýza výrazu „a“ v češtině a angličtině). Pokud není jazyk identifikován v dokumentu ALTO, můžeme údaj o jazyce získat z metadat o publikaci, ale stává se také, že jedna publikace obsahuje pasáže v několika jazycích, což se může sice objevit v bibliografických datech, ale už nebývá uvedeno, na jakých stránkách se dané jazyky vyskytují. Existují také případy, kdy je cizojazyčná pasáž (kapitola, odstavec) nebo pouze jednotlivá slova součástí hlavního textu psaného jedním jazykem.[38] Při špatné kombinaci vstupního textu a rozpoznávaného jazyka dochází k nekvalitnímu výstupu (jednotlivá slova jsou chybně označena za lemmata aj.). Tomu by bylo možné předejít vstupní analýzou jazyka pasáže (odstavce) odesílané externím službám ke zpracování, např. pomocí aplikace Apache Tika.[39]

4 DESIDERATA

Jednou z překážek pro snadné propojení textu s jeho umístěním na stránce je způsob, jakým nástroje pro OCR segmentují rozpoznaný text. Ve výstupech ALTO se interpunkce objevuje jako součást jednoho elementu **<String>**, zatímco aplikace NLP často považují interpunkci za samostatný token. Tím dochází k nesouladu mezi počtem elementů pro text v dokumentu ALTO a mezi elementy TEI, které vzniknou na základě analýzy nástroji pro zpracování přirozeného jazyka. To vede ke složitějšímu propojení mezi jednotlivými formáty (tj. textovým obsahem a jeho umístěním na stránce).

Pokud se podaří zapojit lepší detekci struktury textu, bude možné realizovat např. rozčlenění veršovaných textů na básně, sloky a jednotlivé verše, identifikovat seznamy, tabulky apod.

Až budou k dispozici ve formátu ALTO další elementy, např. o toku textu mezi jednotlivými bloky textu nebo o rozpoznaných entitách, bude je možné využít při extrakci textu pro externí služby nebo rovnou začlenit do výstupu TEI ve formě odpovídajících elementů.

Rovněž lze vylepšit zpracování rozpoznaných entit vytvořením seznamu s jedinečnými údaji o identifikovaných osobách, místech a organizacích (**<listPerson>**, **<listPlace>**, **<listOrg>**), které lze doplnit o údaje z autoritních databází nebo Wikidat.[40]

Současné implementace využívají pouze dvě externí služby, pokud se podaří zajistit konverzi výstupů do TEI z jiných/dalších externích služeb, bude je možné integrovat do celého transformačního procesu.

V neposlední řadě by bylo vhodné implementovat podporu pro další typy digitálních knihoven, které využívají odlišné metadatové a datové formáty, jako jsou např. IIIF (pro popis digitalizovaných publikací) nebo hOCR (pro zachycení textu).

ZÁVĚR

Digitální knihovny obsahují velké množství písemného kulturního dědictví i vědeckých poznatků našich předků i současníků. Jejich zpřístupnění v podobě, která odpovídá standardu TEI P5 a obsahuje navíc jazykovou analýzu textu, významnou měrou přispěje ke studiu a další analýze těchto zdrojů moderními metodami digitálních humanitních věd. Rovněž se tím podpoří rozvoj strojového učení a vznik dalších a kvalitnějších velkých jazykových modelů. Existující implementace konverze digitálních publikací a jejich metadat do formátu TEI P5 ukázaly, že se současnými nástroji a službami lze dosáhnout automatizovaného zpracování repozitářů digitálních knihoven, i když je zde prostor pro další zlepšení kvality výstupu. Pro další badatelskou práci s dokumenty je výhodné mít k dispozici všechny potřebné údaje i samotný text zachycené pomocí jediného standardu. Rovněž tím vniká prostor pro vytváření obecných nebo specializovaných korpusů ve formátu, který umožňuje zachovávat vazbu na digitální předlohu nejen na úrovni celku, ale i stran a dalších detailnějších úseků.

DEDIKACE

Článek vznikl na základě institucionální podpory dlouhodobého koncepčního rozvoje výzkumné organizace poskytované Ministerstvem kultury ČR.

■ **Recenzovaný článek / Reviewed article**

ZDROJE

BERLIN-BRANDENBURGISCHE AKADEMIE DER WISSENSCHAFTEN, (2007–2025). *Deutsches Textarchiv*. Online. Berlin: Berlin-Brandenburgische Akademie der Wissenschaften. Dostupné z: https://www.deutschestextarchiv.de. [cit. 2025-11-11].

BURNARD, Lou; SCHÖCH, Christof a Carolin ODEBRECHT, (2021). *In search of comity: TEI for distant reading*. Online. Journal of the Text Encoding Initiative. 2021-03-17, Issue 14. ISSN 2162-5603. Dostupné z: https://doi.org/10.4000/jtei.3500. [cit. 2025-12-17].

DALMAU, Michelle a Kevin HAWKINS, (2014). *"Reports of My Death Are Greatly Exaggerated": Findings from the TEI in Libraries Survey*. Online. Journal of the Text Encoding Initiative. Issue 8. ISSN 2162-5603. Dostupné z: https://doi.org/10.4000/jtei.1322. [cit. 2025-12-17].

EarlyPrint Library, ([bez data]). Online. Dostupné z: https://texts.earlyprint.org/exist/apps/shc/home.html. [cit. 2025-11-11].

Handwriting recognition, (2001-). Online. In: Wikipedia: the free encyclopedia. San Francisco (CA): Wikimedia Foundation, 3 October 2025. Dostupné z: https://en.wikipedia.org/wiki/Handwriting_recognition. [cit. 2025-11-11].

HUTAŘ, Jan; ŠVÁSTOVÁ, Pavla; KVASNICA, Jaroslav; LODROVÁ, Iveta; PAVČÍK, Filip et al., (2024). *OCR (ALTO XML a TXT OCR): Předpis pro zápis ALTO XML pro užití v rámci Standardu NDK*. Online. Verze 1.0. Dostupné z: https://standardy.ndk.cz/ndk/standardy-digitalizace/ALTO_XML_final_1.0.pdf. [cit. 2025-11-11].

CHIFFOLEAU, Floriane, (2024). *Keeping It Open: A TEI-based Publication Pipeline for Historical Documents*. Online. Journal of the Text Encoding Initiative. Issue 15. ISSN 2162-5603. Dostupné z: https://doi.org/10.4000/12s01. [cit. 2025-12-17].

ISTEX: Infrastructure de services pour la fouille de textes, ([bez data]). Online. Dostupné z: https://www.istex.fr. [cit. 2025-11-11].

LEHEČKA, Boris, (2023). *Standardy pro zachycení výsledků rozpoznání textu*. Online. ITlib. Informačné technológie a knižnice. 2023-12-01, roč. 27, č. SC2, s. 63-74. ISSN 1335-793X. Dostupné z: https://doi.org/10.52036/1335793x.2023.sc2.63-74. [cit. 2025-11-11].

LEHEČKA, Boris, (2025a). *Converting data from Czech digital libraries to TEI: TEI Annual Conference and Members' Meeting 2025 (TEI 2025)*. Version v1. Krakau. Dostupné také z: https://doi.org/10.5281/zenodo.17216539.

LEHEČKA, Boris, (2025b). *Enriching textual data from digital libraries.: TEI Annual Conference and Members' Meeting 2025 (TEI 2025)*. Version v1. Krakau. Dostupné také z: https://doi.org/10.5281/zenodo.17216539.

LEHEČKA, Dalibor; NOVÁK, David; KERSCH, Filip; HLADÍK, Roman; BÍŠKOVÁ, Jarmila et al., (2022). Certifikovaná metodika, Osvědčení č. 256. *Metodika přípravy dat z digitálních knihoven pro využití v digitálních humanitních vědách*. Dostupné také z: http://www.nusl.cz/ntk/nusl-511549.

LHOTÁK, Martin, (2020). *DL4DH – Digital Libraries for Digital Humanities: nový projekt na vytěžování obsahu digitálních knihoven*. Online. ITlib. Informačné technológie a knižnice. Roč. 2020, č. 4, s. 26–31. ISSN 1335-793X. Dostupné z: https://itlib.cvtisr.sk/wp-content/uploads/2021/02/Lhotak.pdf. [cit. 2025-12-15].

LIBRARY OF CONGRESS, ([bez data]). *ALTO: Technical metadata for Layout and Text Objects*. Online. Dostupné z: https://www.loc.gov/standards/alto/. [cit. 2025-11-11].

LIBRARY OF CONGRESS, (2024). *MODS: Metadata Object Description Schema*. Online. Dostupné z: https://www.loc.gov/standards/mods/. [cit. 2025-11-11].

NÁRODNÍ KNIHOVNA ČR, (2024). *Standardy pro metadata: Národní digitální knihovna*. Online. NÁRODNÍ KNIHOVNA ČR. Národní digitální knihovna. 17. 12. 2024. Dostupné z: https://standardy.ndk.cz/ndk/standardy-digitalizace. [cit. 2025-11-11].

Optical character recognition, (2001-). Online. In: Wikipedia: the free encyclopedia. San Francisco (CA): Wikimedia Foundation, 3 November 2025. Dostupné z: https://en.wikipedia.org/wiki/Optical_character_recognition. [cit. 2025-11-11].

Registr Krameriů, ([bez data]). Online. Brno: Moravská zemská knihovna v Brně. Dostupné z: https://registr.digitalniknihovna.cz. [cit. 2025-12-16].

SCHEITHAUER, Hugo; CHAGUÉ, Alix a Laurent ROMARY, (2024). *Which TEI Representation for the Output of Automatic Transcriptions and Their Metadata? An Illustrated Proposition*. Online. Journal of the Text Encoding Initiative. Issue 15. ISSN 2162-5603. Dostupné z: https://doi.org/10.4000/13e5a. [cit. 2025-12-17].

STRAKA, Milan a Jana STRAKOVÁ, (2014). *NameTag*. Online. Praha: LINDAT/CLARIAH-CZ, digitální knihovna při Ústavu formální a aplikované lingvistiky, Matematicko-fyzikální fakulta Univerzity Karlovy. Dostupné z: http://hdl.handle.net/11858/00-097C-0000-0023-43CE-E. [cit. 2025-11-11].

STRAKA, Milan a Jana STRAKOVÁ, (2016). *UDPipe*. Online. Praha: LINDAT/CLARIAH-CZ, digitální knihovna při Ústavu formální a aplikované lingvistiky, Matematicko-fyzikální fakulta Univerzity Karlovy. Dostupné z: http://hdl.handle.net/11234/1-1702. [cit. 2025-11-11].

TEI CONSORTIUM, (September 2018). *Best Practices for TEI in Libraries: A guide for mass digitization, automated workflows, and promotion of interoperability with XML using the TEI*. Online. TEI CONSORTIUM. Text Encoding Initiative. Dostupné z: https://www.tei-c.org/extra/teiinlibraries/. [cit. 2025-11-11].

Text Creation Partnership, ([bez data]). Online. Ann Arbor: Text Creation Partnership. Dostupné z: http://www.textcreationpartnership.org. [cit. 2025-11-11].

POZNÁMKY

[1] Viz https://github.com/moravianlibrary/libri-augmentati.

[2] Viz https://github.com/LIBCAS/DL4DH-TEI-Converter.

[3] V případě potřeby je možné do stejného dokumentu zahrnout relevantní data XML z jiného jmenného prostoru, např. MODS nebo DublinCore, a to v elementu **<xenoData>**.

[4] Viz https://escriptorium.rich.ru.nl.

[5] Viz https://www.transkribus.org.

[6] Viz např. https://github.com/cneud/ocr-conversion, https://github.com/UB-Mannheim/ocr-fileformat nebo https://github.com/FloChiff/workshop-discholed-tei2022/.

[7] Viz např. https://github.com/kat-kel/alto2tei.

[8] Viz https://leaf-writer.leaf-vre.org.

[9] https://gitlab.com/maartenes/TEITOK

[10] Tokenizace (rozdělení na tokeny) probíhá na základě mezery mezi slovy, případně vybraných interpunkčních znaků.

[11] Viz https://teipublisher.com/exist/apps/tei-publisher-home/index.html. Jedná se o soustavu několika relativně samostatných knihoven, jejichž zdrojový kód je k dispozici v repozitáři https://github.com/eeditiones/.

[12] https://wiki.korpus.cz/doku.php/pojmy:korpus

[13] https://wiki.korpus.cz/doku.php/:pojmy:struktura_korpusu

[14] Obsahuje digitální edice téměř 60 tisíc anglických knih.

[15] Obsahuje přes 30 milionů dokumentů od velkých mezinárodních vydavatelství (Elsevier, Springer, Wiley apod.) i z volně přístupných úložišť (PLoS, SciELO aj.).

[16] Až budou v české vědě při hodnocení výstupů digitální vědecké edice postaveny na roveň edicím tištěným, můžeme očekávat rostoucí zájem o standard TEI i u nás.

[17] Podrobněji viz např. Lehečka et al. (2022, s. 22–24) a úložiště České expedice (https://github.com/ceskaexpedice).

[18] https://wiki.lyrasis.org/pages/viewpage.action?pageId=66585857

[19] https://www.loc.gov/standards/mods/

[20] https://www.dublincore.org

[21] Podrobnější popis viz (Lehečka 2023).

[22] Pro tyto elementy je potřeba podpora na straně softwaru pro OCR a aktuálně žádné dokumenty v digitálních knihovnách, s kterými jsme pracovali, tyto elementy neobsahují. S jejich generováním a využitím se počítá v aktuálně probíhajících výzkumných projektech, jako je např. „Orbis Pictus – oživení knihy pro kulturní a kreativní odvětví“ (projekt č. DH23P03OVV033, podpořený Ministerstvem kultury ČR). Až budou taková data k dispozici, výrazně to vylepší informace o struktuře a obsahu stránky i ve formátu TEI.

[23] Element **<Styles>** je nepovinný, takže nástroje OCR nemusejí typografické kvality textu zachycovat (a DMF je nevyžaduje).

[24] S atributy **<String>**/@SUBS_TYPE, @WC, @CC a @CS ani s elementy **<ALTERNATIVE>**, **<Glyph>** a **<Variant>** jsme se v testovacích datech zatím nesetkali.

[25] Viz https://www.rfc-editor.org/info/bcp47.

[26] Viz např. https://ocr-d.de/en/gt-guidelines/trans/trSilbentrennung.html.

[27] Zdrojový kód viz https://github.com/ufal/udpipe.

[28] Zdrojový kód viz https://github.com/ufal/nametag.

[29] Nechtěli jsme některý z formátů upřednostnit, ale nabídnout pokud možno stejný základ pro badatele s různými zájmy a programovým vybavením. Cílem tedy nebylo generovat kvalitní kritické edice digitalizovaných publikací nebo vytvářet prostředí, které bude ke vzniku takových edic sloužit. Nicméně příklon ke standardu TEI, který k takovým účelům slouží, může vznik kritických edic v tomto formátu podpořit.

[30] Ke konci roku 2025 jde zhruba o 293 tis. monografií, 7476 periodik, tj. přibližně 149,5 tis. stran, z nichž je 27,5 tis. veřejně dostupných. Čísla vycházejí z údajů o České digitální knihovně (https://ceskadigitalniknihovna.cz), která obsahuje metadata průběžně sklízená z digitálních knihoven provozovaných v České republice. Údaje se průběžně aktualizují a zobrazují na webu (Registr Krameriů [bez data]).

[31] Viz https://github.com/LIBCAS/DL4DH-Kramerius-plus.

[32] Viz https://github.com/LIBCAS/DL4DH-TEI-Converter.

[33] V současnosti běží pouze jedna instance aplikace DL4DH Feeder pro přístup ke sbírkám Národní knihovny ČR (https://feeder.nkp.cz). Další instalace se testují v Knihovně akademie věd ČR a v Moravské zemské knihovně v Brně.

[34] Viz např. https://github.com/moravianlibrary/libri-augmentati/blob/main/src/xslt/conversion/mods-to-teiHeader.xsl.

[35] Postup zatím nebyl implementován do existujících nástrojů.

[36] Viz https://wiki.korpus.cz/doku.php/pojmy:token.

[37] Viz https://universaldependencies.org/docs/format.html.

[38] Tady se ukazuje nevýhoda nástrojů NLP, které primárně pracují s prostým textem. Strukturovaný formát, který by obsahoval např. i údaj o (měnícím se) jazyce by mohl zlepšit analýzu vícejazyčných textů.

[39] Viz https://tika.apache.org.

[40] Viz https://www.wikidata.org.

(Nad)typ	Popis	Příklad	TEI
a	ČÍSLA JAKO SOUČÁSTI ADRES		
ah	číslo popisné	ul. Šikmá `<ne type="ah">`356`</ne>`	`<tei:num><tei:w>`356`</tei:w></tei:num>`
at	telefon, fax	tel. `<ne type="at">`256 458 588`</ne>`	`<tei:num type="phone"><tei:w>`256`</tei:w><tei:w>`458`</tei:w><tei:w>`588`</tei:w></tei:num>`
az	PSČ	`<ne type="az">`180 00`</ne>` Praha 8	`<tei:num type="zip"><tei:w>`180`</tei:w><tei:w>`00`</tei:w></tei:num>`
C	Komplexní bibliografický údaj	`<ne type="ne">`...`</ne>`	`<tei:bibl>` (pokud jde o jediný prvek v rámci `<tei:s>`)
C	Komplexní bibliografický údaj	`<ne type="ne">`...`</ne>`	`<tei:objectName type="bibliography" ana="#nametag-C">` (pokud nejde o jediný prvek v rámci `<tei:s>`)
g	GEOGRAFICKÉ NÁZVY		
gc	státní útvary	`<ne type="gc">`Česká republika`</ne>`, `<ne type="gc">`Svatá říše římská`</ne>`	`<tei:placeName><tei:country><tei:w>`Česká`</tei:w><tei:w>`republika`</tei:w></tei:country></tei:placeName>`

(Nad)typ	Popis	Příklad	TEI
gh	vodní útvary	`<ne type="gh">`Vltava`</ne>`, `<ne type="gh">`Balaton`</ne>`	`<tei:geogName type="water"><tei:w>`Vltava`</tei:w></tei:geogName>`
gl	přírodní oblasti / útvary	`<ne type="gl">`Sibiř`</ne>`, `<ne type="gl">`Apeninský poloostrov`</ne>`, `<ne type="gl">`Polabí`</ne>`	`<tei:geogName type="area"><tei:w>`Apeninský`</tei:w><tei:w>`poloostrov`</tei:w></tei:geogName>`
gq	části obcí, pomístní názvy	`<ne type="gq">`Smíchov`</ne>`	`<tei:placeName><tei:settlement><tei:w>`Smíchov`</tei:w></tei:settlement></tei:placeName>`
gr	menší územní jednotky	`<ne type="gr">`Morava`</ne>`, `<ne type="gr">`Rychnovsko`</ne>`, `<ne type="gr">`Badensko-Württembersko`</ne>`	`<tei:placeName><tei:region><tei:w>`Morava`</tei:w></tei:region></tei:placeName>`
gs	ulice, náměstí	`<ne type="gs">`ul. Šikmá`</ne>`, `<ne type="gs">`nám. Míru`</ne>`	`<tei:address><tei:street><tei:w>`ul.`</tei:w><tei:w>`Šikmá`</tei:w></tei:street></tei:address>`
gt	kontinenty	`<ne type="gt">`Jižní Amerika`</ne>`	`<tei:geogName type="continent"><tei:w>`Jižní`</tei:w><tei:w>`Amerika`</tei:w></tei:geogName>`
gu	obce, hrady a zámky	`<ne type="gu">`Praha`</ne>`, `<ne type="gu">`Opočno`</ne>`	`<tei:placeName><tei:settlement><tei:w>`Praha`</tei:w></tei:settlement></tei:placeName>`
g_	geografický název nespecifikovaného typu / nezařaditelný do ostatních typů	Učkuduk v `<ne type="g">`Mojunkumech`</ne>`	`<tei:placeName><tei:w>`Mojunkumech`</tei:w></tei:placeName>`

(Nad)typ	Popis	Příklad	TEI
i	NÁZVY INSTITUCÍ		
ia	přednášky, konference, soutěže,...	`<ne type="ia">`Stanley Cup`</ne>`	`<tei:objectName><tei:w>`Stanley`</tei:w><tei:w>`Cup`</tei:w></tei:objectName>`
ic	kulturní, vzdělávací a vědecké instituce, sportovní kluby,...	kino `<ne type="ic">`Dlabačov`</ne>`, hokejisté `<ne type="ic">`Sparty`</ne>`	`<tei:orgName><tei:w>`Dlabačov`</tei:w></tei:orgName>`
if	firmy, koncerny, hotely,...	`<ne type="if">`Unipetrol`</ne>`, řetězec `<ne type="if">`Edeka`</ne>`	`<tei:orgName><tei:w>`Unipetrol`</tei:w></tei:orgName>`
io	státní a mezinárodní instituce, politické strany a hnutí, náboženské skupiny	`<ne type="io">`Evropská komise`</ne>`, `<ne type="io">`Policie`</ne>`, `<ne type="io">`ODS`</ne>`	`<tei:orgName><tei:w>`Policie`</tei:w></tei:orgName>`
i_	instituce nespecifikovaného typu / nezařaditelné do ostatních typů	`<ne type="i">`Studijní odd.`</ne>`	`<tei:orgName><tei:w>`Studijní`</tei:w><tei:abbr><tei:w>`odd`</tei:w><tei:pc>`.`</tei:pc></tei:abbr></tei:orgName>`
m	NÁZVY MÉDIÍ		

(Nad)typ	Popis	Příklad	TEI
me	e-mailové adresy	<ne type="me">john@seznam.cz</ne>	<tei:email>john@seznam.cz</tei:email>
mi	internetové odkazy	<ne type="mi">www.cinestar.cz</ne>	<tei:ref target="www.cinestar.cz">www.cinestar.cz</tei:ref>
mn	periodika, redakce, tiskové agentury	agentura <ne type="mn">Interfax</ne>	<tei:orgName><tei:w>Interfax</tei:w></tei:orgName>
ms	rozhlasové a televizní stanice	<ne type="ms">Frekvence 1</ne>, <ne type="ms">ČT 2</ne>	<tei:orgName><tei:w>ČT</tei:w><tei:w>2</tei:w></tei:orgName>
n	ČÍSLA SE SPECIFICKÝM VÝZNAMEM		
na	věk	je mu <ne type="na">18</ne>	<tei:num><tei:w>18</tei:w></tei:num>
nc	číslo s významem počtu	s <ne type="nc">35miliardovým</ne> deficitem, od <ne type="nc">8</ne> do <ne type="nc">40</ne>	<tei:num><tei:w>35miliardovým</tei:w></tei:num>
nb	číslo strany, kapitoly, oddílu, obrázku	podle paragrafu <ne type="nb">187</ne>, tab na s. <ne type="nb">13</ne>	<tei:num><tei:w>187</tei:w></tei:num>
no	číslo s významem pořadí	<ne type="no">154.</ne> vydání, umění <ne type="no">20.</ne> století	<tei:num type="ordinal"><tei:w>20.</tei:w></tei:num>
ns	sportovní skóre	vyhráli <ne type="ns">5:0</ne>	<tei:num>5</tei:num><tei:pc>:</tei:pc><tei:num>0</tei:num>

(Nad)typ	Popis	Příklad	TEI
ni	itemizátor	`<ne type="ni">`1)`</ne>` DPH, `<ne type="ni">`2)`</ne>` daň z příjmu	`<tei:num><tei:w>`1`</tei:w><tei:pc>`)`</tei:pc></tei:num>`
n_	číslo se specifickým významem, jehož typ nebyl vyčleněn jako samostatný / nelze identifikovat	autobusová linka `<ne type="n">`158`</ne>`	`<tei:num>`158`</tei:num>`
o	NÁZVY VĚCÍ		
oa	kulturní artefakty (knihy, filmy stavby,...)	Hugovi `<ne type="oa">`Bídníci`</ne>`	`<tei:objectName type="artefact"><tei:w>`Bídníci`</tei:w></tei:objectName>`
oe	měrné jednotky (zapsané zkratkou)	200 `<ne type="om">`MHz`</ne>`	`<tei:unit><tei:w>`MHz`</tei:w></tei:unit>`
om	měny (zapsané zkratkou, symbolem)	1000 `<ne type="om">`Kč`</ne>`, 30 `<ne type="om">`$`</ne>`	`<tei:unit><tei:w>`Kč`</tei:w></tei:unit>`
op	výrobky	mikroprocesory `<ne type="op">`Intel Pentium`</ne>`	`<tei:objectName type="product"><tei:w>`Intel`</tei:w><tei:w>`Pentium`</tei:w></tei:objectName>`

(Nad)typ	Popis	Příklad	TEI
or	předpisy, normy,..., jejich sbírky	ve <ne type="or">Sbírce zákonů</ne>	<tei:objectName type="rule"><tei:w>Sbírce</tei:w><tei:w>zákonů</tei:w></tei:objectName>
o_	názvy nespecifikovaných o typu / nezařaditelné do ostatních typů	čeleď <ne type="o">Rubiaceae</ne>, <ne type="o">HIV</ne>	<tei:objectName><tei:w>HIV</tei:w></tei:objectName>
p	JMÉNA OSOB		
pc	obyvatelská jména	<ne type="pc">Afričan</ne>, <ne type="pc">Pražan</ne>	<tei:name type="population"><tei:w>Afričan</tei:w></tei:name>
pd	titul (pouze zkratkou)	<ne type="pd">Mgr.</ne> J. Pola	<tei:abbr><tei:w>Mgr</tei:w><tei:pc>.</tei:pc></tei:abbr>
pf	křestní jméno	<ne type="pf">David</ne> Uhl, <ne type="pf">J.</ne> Drda	<tei:forename><tei:w>David</tei:w></tei:forename>, <tei:forename><tei:w>J</tei:w><tei:pc>.</tei:pc></tei:forename>
pm	druhé křestní jméno	Georg <ne type="pm">Friedrich</ne> Händel	<tei:forename type="middle"><tei:w>Friedrich</tei:w></tei:forename>
pp	náboženské postavy, pohádkové a mytické postavy, personifikované vlastnosti	<ne type="pp">sv. Jakub</ne>, <ne type="pp">Prozřetelnost</ne>	<tei:persName><tei:w>sv</tei:w><tei:pc>.</tei:pc><tei:w>Jakub</tei:w></tei:persName>

(Nad)typ	Popis	Příklad	TEI
ps	příjmení	<ne type="ps">Nováková</ne>, <ne type="ps">van Dyk</ne>	<tei:surname>Nováková</tei:surname>
p_	jméno osoby nespecifikovaného typu / nezařaditelné do ostatních typů	<ne type="p">Slované</ne>	<tei:name><tei:w>Slované</tei:w></tei:name>
t	ČASOVÉ ÚDAJE		
td	den	<ne type="td">1.</ne> října 2005	<tei:date><tei:w>1</tei:w><tei:pc>.</tei:pc></tei:date>
tf	svátky a významné dny	<ne type="tf">Velikonoce</ne>	<tei:date type="holiday"><tei:w>Velikonoce</tei:w></tei:date>
th	hodina	v <ne type="th">17</ne> hodin	<tei:time><tei:w>17</tei:w></tei:time>
tm	měsíc	1. <ne type="tm">října</ne> 2005	<tei:date><tei:w>října</tei:w></tei:date>
ty	rok	od roku <ne type="ty">1555</ne>	<tei:date><tei:w>1555</tei:w></tei:date>

Tabulka 1 Typy entit rozpoznávané nástrojem NameTag a odpovídající elementy TEI

MODS	TEI
<mods:modsCollection>	<tei:teiHeader>
<mods:mods>	<tei:fileDesc>/<tei:titleStmt>
<mods:titleInfo>	<tei:title>
<mods:nonSort>, <mods:title>	textový obsah elementu <tei:title>
<mods:subTitle>	<tei:title type='sub'>

MODS	TEI
<mods:name>[**mods:role/mods:roleTerm**[@authority = 'marcrelator' and @type = 'code'][. = ('aut', 'aui')]]	**<tei:author>**
<mods:namePart>[not(@type)]	**<tei:persName>**
<mods:nameIdentifier>	**<tei:idno>**
<mods:name>[**mods:role/mods:roleTerm**[@authority = 'marcrelator' and @type = 'code'][. = ('com')]]]	**<tei:editor>**
<mods:publisher>	**<tei:publisher>**
<mods:placeTerm>[not(@authority)][parent::mods:place]	**<tei:pubPlace>**
<mods:dateIssued>	**<tei:date>**
<mods:languageTerm>	**<tei:language>**
<mods:extent>[parent::mods:physicalDescription]	**<tei:extent>**
Prvky, které se nezpracovávají	
<mods:typeOfResource>	
<mods:genre>	
<mods:placeTerm>[@authority = 'marccountry' and @type = ('code', 'text')]	
<mods:issuance>	

Tabulka 2 Převod mezi elementy standardu MODS a TEI

Pozn.: v hranatých závorkách je uvedena podmínka vyjádřená v jazyce XPath, kterou musí daný element MODS splňovat.