SMART DIGILINKA – STROJOVÉ UČENÍ PRO DIGITALIZACI TIŠTĚNÉHO DĚDICTVÍ

Mgr. Jana Krýdová; krydova@knav.cz; (Knihovna AV ČR, v. v. i.)
Ing. Jan Kohút; ikohut@fit.vut.cz; (Vysoké učení technické v Brně)
Ing. Petr Žabička; petr.zabicka@mzk.cz; (Moravská zemská knihovna v Brně)
Ing. Michal Hradiš; PhD.; ihradis@fit.vut.cz; (Vysoké učení technické v Brně)
Ing. Martin Lhoták; lhotak@knav.cz; (Knihovna AV ČR, v. v. i.)
Ing. Petr Kukač; petr.kukac@nkp.cz; (Národní knihovna ČR)

Účel – Cílem projektu „Smart digilinka – strojové učení pro digitalizaci tištěného dědictví" je vytvoření nástrojů pro automatizovanou tvorbu metadat digitalizovaných dokumentů – zejména knih a periodik. Článek popisuje tento projekt a jeho výstupy a také stručně popisuje vznik popisných metadat v procesu digitalizace knihovních fondů.

Metoda – Metadata budou z obrazových souborů extrahována pomocí nástrojů strojového učení a výsledky budou upravovány prostřednictvím zpětné vazby, kterou poskytnou uživatelé, což jsou v tomto případě pracovníci knihoven, kteří vytvářejí metadata k digitalizovaným dokumentům.

Výsledky – Výsledky projektu budou tři nástroje a dvě metodiky. Nástroj Exposea slouží k automatickému vyrovnání předloh, což usnadní manipulaci s velkými, fyzicky poškozenými předlohami. Nástroj Metakat umožňuje vytvářet automatizované metadatové popisy jednotlivých stran digitalizovaných dokumentů a také analytických úrovní dokumentu (kapitoly pro monografie a články pro periodika). Pro trénování systému automatické extrakce metadat byly vytvořeny sady ručně anotovaných stran a kapitol, na nichž by se mohl systém učit extrahovat metadata. Navazující nástroj Metacheck dokáže vytvářet a odesílat zpětnou vazbu pro zlepšení kvality vytvořených metadat. Dalšími dvěma výstupy projektu jsou Metodika scelování po částech skenovaných předloh a Metodika kontroly kvality v digitalizační lince. Celkovým výsledkem je Smart Digilinka, která představuje kombinaci všech nástrojů, implementovanou formou poloprovozu na digitalizačním pracovišti, které využívá systém ProArc.

Originalita/hodnota – Projekt Smart Digilinka umožní významně zrychlit a obohatit proces vytváření metadat k digitalizovaným dokumentům, čímž se také zkrátí doba jejich digitalizace a doba, za kterou se mohou dostat k uživatelům.

http://doi.org/10.52036/1335793X.2024.2.32-36

ÚVOD

V českých knihovnách se stále nachází velké množství dokumentů, které jsou ohroženy degradací materiálů, na nichž jsou vytištěny. Obsah takových dokumentů je možné zachovat jejich digitalizací, která však není dostatečně rychlá. Což má za následek, že tyto poškozené knihy a časopisy mohou být uživatelům ve fyzické podobě znepřístupněny, a protože ještě nejsou zdigitalizovány, uživatel se k nim nedostane ani digitálně. Projekt s názvem „Smart Digilinka – strojové učení pro digitalizaci tištěného dědictví" má za cíl zefektivnit a zrychlit digitalizaci fondů knihoven prostřednictvím vytvoření poloprovozu digitalizační linky, která bude využívat nástroje strojového učení přímo v procesu digitalizace (*Smart Digilinka* 2024*)*.

Projekt je plánován na roky 2023 až 2027 a v jeho průběhu vznikne několik nástrojů a dvě metodiky, které umožní časově a personálně nejnáročnější procesy digitalizace zrychlit a zároveň vylepšit její výstupy, zejména co se týče metadat. Výsledky projektu jsou kombinací již existujících nástrojů a nástrojů, které jsou vyvíjeny přímo pro účely projektu. Tým je složen z pracovníků Národní knihovny České republiky, Moravské zemské

knihovny v Brně, Knihovny Akademie věd ČR a Vysokého učení technického v Brně. Výsledky projektu budou implementovatelné do nástrojů pro vytváření metadat, které používají zapojené knihovny, tedy do systému ProArc (*ProArc* 2015–) a také do digitalizační linky NK ČR.

VYTVÁŘENÍ METADAT V PROCESU DIGITALIZACE

Knihovny zapojené v projektu Smart digilinka se při digitalizaci řídí definicemi metadatových formátů, které vytváří Oddělení pro standardy NK ČR[1]. Tyto standardy umožňují popis digitalizovaných dokumentů až na analytickou úroveň. Problémem je, že takto podrobný popis mnoho knihoven v praxi nepoužívá, protože trvá mnohem déle než vytvoření základních metadat. Existují Definice metadatových formátů pro monografie (FREMROVÁ et al. 2023), periodika (KOČIŠOVÁ et al. 2023), zvukové dokumenty[2], elektronické publikace[3] a pro staré tisky je vytvořen metodický pokyn[4]. Tyto dokumenty definují nejen popisná (MODS[5], DC[6]), ale také strukturální (METS[7]), technická (PREMIS[8], MIX[9]), administrativní (PREMIS) a autorsko-právní metadata (copyrightMD[10]) digitalizovaných dokumentů. Smart digilinka řeší zejména metadata strukturální a popisná, tedy ta, která vznikají v převážné míře ručně. V současné době vznikají popisná metadata k digitalizovaným dokumentům víceméně na dvou úrovních, na úrovni celého dokumentu a na úrovni jednotlivé strany. V případě periodik jde o úrovně titulu periodika, ročníku, čísla a strany, u monografií jde o titul a stranu.

Kromě Definice metadatových formátů existují ještě dva materiály, které pomáhají při vytváření metadat digitalizovaných dokumentů. Jde o Pravidla popisu monografií (KOČIŠOVÁ a BEŇAČKOVÁ 2023a) a Pravidla popisu periodik (KOČIŠOVÁ a BEŇAČKOVÁ 2023b). Oba tyto dokumenty určují, jak v praxi popisovat digitalizované knihy a periodika a uvádějí také příklady, kdy například plnění konkrétních elementů nemusí být z Definice metadatových formátů zcela jasné.

V průběhu projektu Smart digilinka vzniknou dva klíčové nástroje pro automatické zpracování a kontrolu metadatových záznamů. Automatickou detekci metadat ze skenů dokumentů zajistí systém Metakat. Návazným nástrojem je Metacheck, v němž dojde ke kontrole extrahovaných metadat a následně také zobrazí uživateli (případně pošle do konkrétního editačního rozhraní přes API) jistotu s jakou je konkrétní údaj správný. Tato pravděpodobnost bude vyjádřena jak procentuálně, tak barevně. Uživatel tato metadata zkontroluje a potvrdí, v případě potřeby opraví a odešle je zpět do nástroje Metacheck, aby se mohl systém dále učit.

Smart digilinka je posledním výstupem projektu, naplánovaným na rok 2027, a integruje všechny vzniklé nástroje do jednoho systému, který bude fungovat automatizovaně a od lidského operátora bude vyžadovat pouze kontrolu kvality a zpětnou vazbu (*Smart Digilinka* 2024*)*. Ještě v průběhu projektu bude Smart Digilinka nasazena jako poloprovoz na některém z digitalizačních pracovišť používajících systém ProArc.

METADATA PRO MONOGRAFIE

Metadata pro monografické dokumenty vznikají stažením bibliografického záznamu z katalogu knihovny. Jde tedy o částečně automatizovaný proces. Ovšem ještě před digitalizací konkrétního dokumentu většinou proběhne jeho rekatalogizace, kterou provádí katalogizátor ručně. Další úroveň popisu vzniká, pokud jde o vícedílnou monografii. V takovém případě je vytvořen záznam pro titul vícedílné monografie a dále záznam pro svazek vícedílné monografie[11], tedy konkrétní díl. Další metadatový záznam může (a nemusí) vzniknout, když má kniha přílohu. Metadata pro jednotlivé strany vznikají na digitalizačních pracovištích částečně automatizovaně, většinou lze například hromadně upravovat čísla i typy stran. Typy stran jsou převážně určovány ručně. Úroveň popisu, která ve většině případů není vytvářena vůbec (a standardy pro digitalizaci to umožňují), je kapitola. Jde totiž o časově i personálně náročný proces, pro který neexistují žádné katalogizační záznamy, které by bylo možné použít. Dokument, členěný na kapitoly, by byl samozřejmě uživatelsky mnohem příjemnější a užitečnější než bez nich (KOČIŠOVÁ a BEŇAČKOVÁ 2023a). Mezi další výhody patří například možnost zobrazit záznamy kapitol v systému Kramerius, což umožňuje také generování citací přímo pro konkrétní kapitolu.

V rámci projektu Smart digilinka vznikne systém Metakat, prostřednictvím něhož bude možné automatizovaně extrahovat metadata ke kapitolám (podkapitolám). Jde o název, případně podnázev kapitoly, číslo kapitoly a také autor (spíše u odborné literatury). Podle aktuální verze standardu pro metadata je možné popsat jednu úroveň kapitol. Jde o případ, kdy jsou všechny kapitoly v knize hierarchicky na jedné úrovni[12], což je typické hlavně pro beletrii. Velká spousta monografií (spíše odborné než beletrie) má však složitější strukturu s několika úrovněmi kapitol[14,15], což je zkomplikováno také různými možnostmi číslování kapitol a také tím, že názvy kapitol (hlavně u starší produkce) mohou být mírně odlišné od znění názvů kapitol v obsahu knihy. V současné době je díky projektu Smart digilinka připravena úprava Definice metadatových formátů pro monogra-

fie, aby bylo možné v metadatovém popisu postihnout složitější strukturu kapitol (FREMROVÁ et al. 2023).

V open-source nástroji Label studio[16] byla vytvořena testovací sada dat pro detekci názvů kapitol. Pracovníci knihoven v tomto nástroji označili přímo na skenech jednotlivých stran, kde přesně se nachází název (kapitoly nebo podkapitoly), její číslo, případně bylo třeba označit, že jde o jiný objekt než je název kapitoly. Také byla vyznačena čísla stran a v naskenovaných obsazích knih byly tyto údaje propojeny.

Nástroj pro automatickou detekci metadat umožní také extrakci metadat, která už jsou k dispozici z bibliografického záznamu v knihovním katalogu (autor knihy, název, podnázev, nakladatelské údaje, apod.). Tyto informace bude možné použít například k ověření správnosti údajů v katalogizačním záznamu.

METADATA PRO PERIODIKA

V případě periodik je možné pro popisná metadata použít pouze katalogizační záznam titulu, další úrovně popisu (záznam ročníku, záznam čísla) jsou vytvářeny ručně. Opět může vzniknout záznam pro přílohu periodika (ročníku, čísla), který je nutno vytvořit ručně. S metadaty stran se to má stejně jako v případě monografií. Také u periodik ve většině případů chybí jedna popisná úroveň, a to je článek. Metadata pro články vznikala pouze izolovaně v rámci konkrétních projektů (například v KNAV) a u již dříve digitalizovaných periodik[17]. Rozhodně nejde o proces, který by probíhal v rámci běžné digitalizace. V KNAV probíhá zpracování článkových metadat, ale pouze u eborn dokumentů[18], tedy u periodik, k nimž existují plné texty v digitální podobě a není nutné je skenovat. I tento metadatový popis vzniká však pouze ručně, případně s pomocí databáze ANL[19], odkud je možné stáhnout záznam článku v případě, že existuje.

Smart digilinka pomocí nástroje Metakat umožní detekovat metadata o ročníku (číslo a datum vydání) a o čísle (číslo a datum vydání). Další možnosti představují například zvláštní název čísla, autor čísla, nebo údaj o tom, zda jde o přílohu periodika (ročníku, čísla). Kromě toho bude možné předpřipravit jednotlivé metadatové záznamy ročníků a čísel, což pomůže ušetřit čas zejména u periodik, která mají v jednom ročníku mnoho čísel (typicky jde o denní tisk). Metakat umožní automatické vytváření metadat pro jednotlivé články, což je v případě ručního zpracování velmi časově náročné. Bude možné extrahovat údaje o názvových informacích článku, autorech, případně klíčová slova, abstrakt a typ článku.

METADATA STRAN

ČÍSLO STRANY

Každá digitalizovaná strana je popsána prostřednictvím dvou údajů – čísel, jedno z nich je index strany a druhé číslo strany. Index strany představuje pořadí strany v reprezentaci, bez ohledu na to, co je na straně fyzicky natištěno. Číslo strany odpovídá tomu, co je vytištěno v předloze. V případě, že není vytištěno nic nebo je v tisku chyba, toto číslo se dopočítá podle kontextu okolních stran. Jakým způsobem se dopočítávají čísla stran je uvedeno v Pravidlech pro popis monografií a periodik (FREMROVÁ et al. 2023).

Z hlediska automatizované extrakce čísel stran může být problematické například to, že číslo strany může být v každé knize natištěno na jiném místě, dále různé průběžné číslování, změna formátu čísel (například strany s přílohami, úvodem, předmluvou apod. jsou často číslovány římskými číslicemi[20]) nebo také fakt, že u některých starých tisků mohou být číslovány odstavce a ne samotné strany.

TYP STRANY

Definice metadatových formátů určuje 16 typů stran, jejichž určení je povinné, pokud se v dokumentu nacházejí. Ostatní typy, které jsou uvedeny v Pravidlech popisu, mohou být použity dobrovolně (KOČIŠOVÁ et al. 2023). V praxi velkých digitalizačních pracovišť to většinou funguje tak, že jsou vytvořeny povinné části metadat a ostatní typy stran se z důvodu časové náročnosti neurčují. Další nejasnosti mohou být mezi tím, jakým způsobem se typ strany určuje. Například ilustrace v monografiích není povinné označit, ale i když se zpracovatel rozhodne ji označit, někdo může označovat jen celostránkové ilustrace, někdo všechny, někdo jen ty, které jsou očíslovány, apod.

Pro vytvoření sady dat, na které by bylo možné trénovat systém určování typů stran, bylo potřeba, aby pracovníci knihoven zkontrolovali správnost popisu určitého počtu již popsaných stran. Byl použit již existující nástroj Annotator (viz. obr. 1), kde byl k dispozici náhled konkrétní strany s minimálním kontextem, tedy s malým náhledem předchozí a následující strany, což by v některých případech mohlo pomoci s určením strany. Také je k dispozici informace o počtu stran, které určované straně předcházejí a které po ní následují. U některých typů stran byl takový kontext nedostačující. Například pro přední a zadní přídeští, titulní stranu apod. Pro některé typy stran je potřeba mít k dispozici širší kontext nebo například odkaz na zdrojový dokument, kde by bylo možné vidět alespoň náhledy (u autorsky chráněných dokumentů) všech stran.

Některé typy stran se v současné praxi v podstatě neoznačují, i když to standard umožňuje. Je to tak kvůli pracnosti, takže se prodlouží čas zpracování daného dokumentu. U těchto typů stran by tedy bylo vhodné, aby byly popisovány automatizovaně. Je však potřeba pro ně nalézt v digitalizovaných dokumentech dostatek příkladů a označit je, aby bylo možné na nich systém trénovat.

Zapojené knihovny používají pro vytváření metadat systém ProArc (KNAV a MZK) a NK používá editační rozhraní digitalizační linky NK. Do obou těchto systémů bude možné posílat a dále využívat výstupy nástrojů Metakat a Metacheck. Metakat dostane skeny jednotlivých stran spolu s OCR a ALTO xml. Výstupem budou extrahovaná metadata, k nimž bude připojena míra jistoty s jakou jsou určena, která bude vyjádřena procentuálně a pro uživatele také barevně pro rychlejší rozlišení. Zpracovatel metadata zkontroluje, případně opraví a prostřednictvím nástroje Metacheck odešle tyto opravy jako zpětnou vazbu, která umožní další učení systému. Metodika kontroly kvality v digitalizační lince bude v podstatě návodem, jak pracovat s nástrojem Metacheck. Poskytne postup, jak vyhodnocovat výsledky nástroje Metakat a jakým způsobem nejlépe poskytnout zpětnou vazbu.

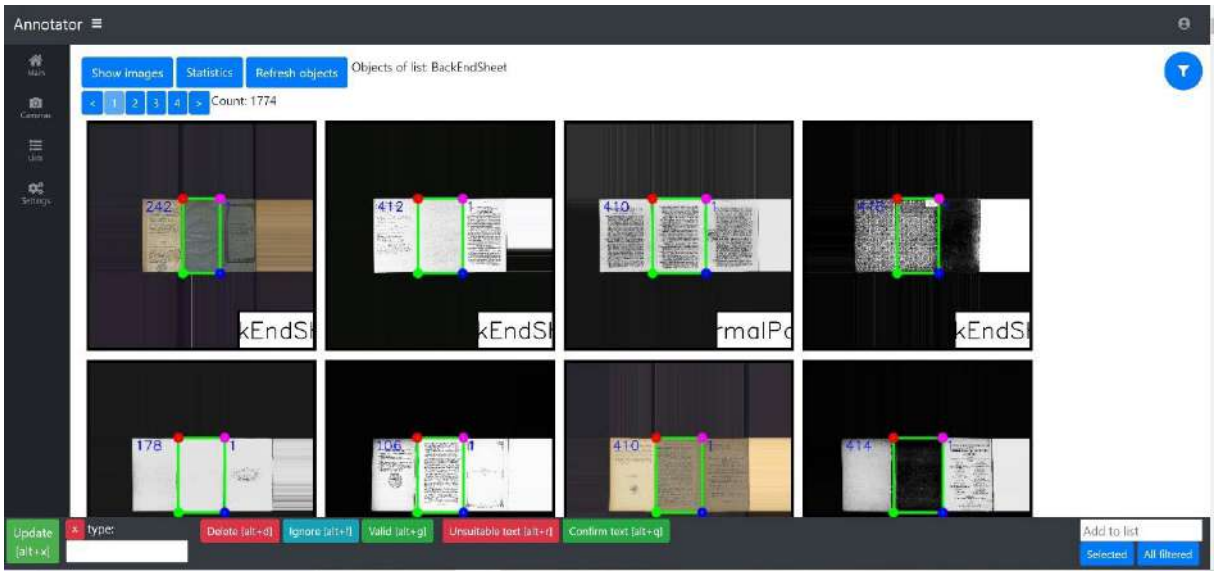
DIGITALIZACE ROZMĚRNÝCH PŘEDLOH

Knihovny mají ve svých fondech řadu dokumentů, jejichž rozměry neumožňují digitalizaci do jednoho souboru. Jde například o rozměrné mapy nebo grafické materiály, plány, apod. Protože nejde o většinovou součást fondů, knihovny kvůli nim nenakupují dostatečně velká skenovací zařízení. V takových případech je možné dokument naskenovat na více částí. Problém pak nastává s tím, jak těchto několik skenů spojit do jednoho souboru, aby mohl být dokument prezentován formou, která bude co nejbližší předloze. To často není možné zejména z důvodu neznalosti vhodných nástrojů. V projektu Smart digilinka vznikne Metodika scelování po částech skenovaných předloh, která poskytne knihovnám a jiným paměťovým institucím doporučení, jak digitalizovat a následně sloučit fyzicky rozměrné dokumenty.

Dále vznikne nástroj Exposea, který bude kombinací již existujících nástrojů, a bude sloužit pro automatické vyrovnání předloh. Systém je zamýšlen pro snímání velmi poškozených předloh (například velkoformátové mapy ve špatném fyzickém stavu) bez nutnosti jejich fyzického narovnání, čímž by mohlo dojít k jejich dalšímu poškození.Tento výsledek je plánován na rok 2025.

ZÁVĚR

Popisná a strukturální metadata vznikají při digitalizaci knihovních fondů stále ve velké míře ruční prací, která je personálně a časově náročná. Projekt Smart digilinka umožní proces vzniku těchto metadat významně urychlit a zefektivnit prostřednictvím jejich automatické extrakce ze skenů digitalizovaných dokumentů. Nástroje pro extrakci metadat – Metakat – a zpětnou vazbu – Metacheck – bude možné implementovat v existujících systémech pro tvorbu metadat. Ještě v průběhu projektu budou vytvořené nástroje nasazeny formou poloprovozu s názvem Smart digilinka do digitalizačního pracoviště využívajícího systém ProArc. Kromě



Obr. 1 Prostředí pro anotování typů stran

nástrojů pro vytváření a kontrolu metadat vzniknou také dvě metodiky – „Metodika pro zcelování po částech skenovaných předloh“ a „Metodika kontroly kvality v digitalizační lince“ – a nástroj s názvem Exposea sloužící k automatickému rovnání předloh.

Dedikace: Projekt "Smart digilinka – strojové učení pro digitalizaci tištěného dědictví", identifikační kód: DH23P03OVV066, je financován z programu NAKI Ministerstva kultury ČR.

LITERATURA

FREMROVÁ, Květa; PAVČÍK, Filip; KOČIŠOVÁ, Pavlína; BEŇAČKOVÁ, Miroslava; OSTRÁKOVÁ, Natalie a další, (2023). *Definice metadatových formátů pro digitalizaci monografických dokumentů*. Online. Dostupné z: https://standardy.ndk.cz/ndk/standardy-digitalizace/monografie21_rev_15_05_23.pdf. [cit. 2024-10-23].

KOČIŠOVÁ, Pavlína a Miroslava BEŇAČKOVÁ, (2023a). *Pravidla pro popis monografií*. Online. Dostupné z: https://standardy.ndk.cz/ndk/standardy-digitalizace/pppmono2.2.verze4.pdf. [cit. 2024-10-23].

KOČIŠOVÁ, Pavlína a Miroslava BEŇAČKOVÁ, (2023b). *Pravidla pro popis periodik*. Online. Dostupné z: https://standardy.ndk.cz/ndk/standardy-digitalizace/copy_of_pppperio8.5_final_15_05_23.pdf. [cit. 2024-10-23].

KOČIŠOVÁ, Pavlína; FREMROVÁ, Květa; JIROUŠEK, Václav; BEŇAČKOVÁ, Miroslava; OSTRÁKOVÁ, Natalie a další, (2023). *Definice metadatových formátů pro digitalizaci periodik*. Online. Dostupné z: https://standardy.ndk.cz/ndk/standardy-digitalizace/DMF_periodika_20_rev_15_05_23.pdf. [cit. 2024-10-23].

ProArc, (2015–). Online. Dostupné z: https://github.com/proarc/. [cit. 2024-10-23].

Smart Digilinka, (2024). Online. Dostupné z: https://smart.lib.cas.cz/. [cit. 2024-10-22].

POZNÁMKY

[1] https://www.nkp.cz/o-knihovne/zakladni-informace/zakladni-dokumenty/vizitky/digi-ochrana3

[2] Pro zvukové dokumenty existuje Definice metadatových formátů pro gramofonové desky: https://standardy.ndk.cz/ndk/standardy-digitalizace/DMF_gramodesky_0.5.pdf a fonografické válečky: https://standardy.ndk.cz/ndk/standardy-digitalizace/DMF_fonovalecky_0.3.pdf

[3] V případě elektronických publikací jde o tři dokumenty, pro elektronické monografie: https://standardy.ndk.cz/ndk/standardy-digitalizace/DMF_ebornmonografie_3.0_final.pdf, elektronická periodika: https://standardy.ndk.cz/ndk/standardy-digitalizace/DMF_ebornperiodika_2.5_final.pdf a skládaná periodika, která kombinují eborn dokumenty s dokumentydigitalizovanými: https://standardy.ndk.cz/ndk/standardy-digitalizace/DMF_eborn_skladana_periodika_1.2.pdf

[4] https://standardy.ndk.cz/ndk/standardy-digitalizace/Metodikapropopisstarchtisk1.pdf

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

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

[7] https://www.loc.gov/standards/mets/

[8] https://www.loc.gov/standards/premis/

[9] https://www.loc.gov/standards/mix/

[10] https://cdlib.org/groups/rights-management-group-copyrightmd/

[11] Příklad vícedílné monografie: https://kramerius.lib.cas.cz/uuid/uuid:562d4075-7b0a-4bc2-9160-02e243875a68

[12] Například zde: https://kramerius.lib.cas.cz/uuid/uuid:9568b3c4-1724-46db-969f-17af2fb37ef9

[13] Příklad knihy, kde jsou použity tři úrovně kapitol: https://kramerius.lib.cas.cz/uuid/uuid:bda10abb-7c1a-46e3-807a-1e760ec763e2

[14] Příklad, kde jsou použity nečíslované podkapitoly: https://kramerius.lib.cas.cz/uuid/uuid:068504a5-e929-4a49-9c4b-df7d59c5b3af

[15] Příklad číslovaných podkapitol: https://kramerius.lib.cas.cz/uuid/uuid:913414c7-6d84-43a5-bd7f-20f82d115c4f

[16] https://labelstud.io/

[17] Příklad skenovaného čísla periodika zpracovaného na úroveň článků a jeho zobrazení v systému Kramerius: https://kramerius.lib.cas.cz/uuid/uuid:d0723154-fb3e-41af-885e-56a2989a88e8

[18] Příklad eborn periodika: https://kramerius.lib.cas.cz/uuid/uuid:58a21319-1b49-45da-858e-af3214e56d88

[19] Databáze Výběr článků v českých novinách, časopisech a sbornících: https://aleph.nkp.cz/F/TQ94AL2CKN59STDYJRP1644BIECDGAN1I5KD762LVUM7VHVHSJ-16828?func=file&file_name=find-b&local_base=ANL

[20] Příklad různého číslování stran v jedné knize: https://kramerius.lib.cas.cz/uuid/uuid:6e465503-7dd1-486c-89af-4b39522bcc9e

■ **Článok bol recenzovaný.**