ORBIS PICTUS: ZPŘÍSTUPNĚNÍ NETEXTOVÝCH DAT Z DIGITÁLNÍCH KNIHOVEN

Mgr. Filip Kersch; kersch@knav.cz; (Knihovna AV ČR, v. v. i.)
Bc. Václav Jiroušek; Vaclav.Jirousek@nkp.cz; (Národní knihovna ČR)
Dalibor Lehečka; Dalibor.Lehecka@mzk.cz; (Moravská zemská knihovna v Brně)
Ing. Martin Kišš; ikiss@fit.vut.cz; (Vysoké učení technické v Brně)
Mgr. Jana Hrzinová; Jana.Hrzinova@nkp.cz; (Národní knihovna ČR)
Ing. Petr Žabička; Petr.Zabicka@mzk.cz; (Moravská zemská knihovna v Brně)
Mgr. Martina Dvořáková; Martina.Dvorakova@mzk.cz; (Moravská zemská knihovna v Brně)
Mgr. Filip Pavčík, PhD.; Filip.Pavcik@nkp.cz; (Národní knihovna ČR)
Ing. Michal Hradiš, PhD.; ihradis@fit.vut.cz; (Vysoké učení technické v Brně)
Mgr. Filip Jebavý; Filip.Jebavy@mzk.cz; (Moravská zemská knihovna v Brně)
Mgr. et Bc. Michaela Bežová; Michaela.Bezova@nkp.cz; (Národní knihovna ČR)
Ing. Martin Lhoták; lhotak@knav.cz; (Knihovna AV ČR, v. v. i.)
Mgr. Květa Fremrová; Kveta.Fremrova@nkp.cz; (Národní knihovna ČR)

Účel – Projekt „Orbis Pictus – oživení knihy pro kulturní a kreativní odvětví" si klade za cíl zpřístupnit netextový obsah českých digitálních knihoven, který je ve srovnání s textovými daty obtížně dosažitelný a neprohledatelný. Tento článek přináší přehled plánovaných výstupů projektu s důrazem na klíčové výsledky dosažené v prvních dvou letech.

Metody – Zpřístupnění netextových objektů v digitalizovaných dokumentech lze rozdělit na tři úlohy: detekci, popis a vyhledání. Identifikaci, lokalizaci a kategorizaci objektů zajistí nástroj AnnoPage, který umožní extrakci popisů objektů a jejich uložení ve standardizovaném formátu. V dalších fázích projektu naváže na AnnoPage nástroj PeopleGator, který identifikuje osoby na fotografiích či kresbách a umožní propojení dokumentů s vyobrazením stejné osoby a vytvoření databáze identifikovaných osob. Projekt bude zakončen softwarovým řešením integrujícím všechny vyvinuté nástroje.

Výsledky – V prvních dvou letech projektu byla vytvořena metodika pro zpracování obrazových dokumentů. Ta popisuje způsob detekce netextových objektů, jejich rozdělení do 25 kategorií a zápis informací pomocí mezinárodních standardů, čímž pokládá základ pro nástroj AnnoPage. K detekci objektů je využíván detektor trénovaný na vlastní datové sadě. Detekované objekty jsou popsány pomocí vektorových reprezentací a textových popisů.

Originalita/hodnota – Výstupy projektu budou integrovány do České digitální knihovny, což umožní využívání vyvinutých nástrojů širokému spektru knihoven, které platforma agreguje. Orbis Pictus je unikátní projekt v oblasti digital humanities díky rozsáhlému shromáždění netextových dat. Výsledky najdou uplatnění nejen v identifikaci objektů a metadat, ale i ve výzkumu a kulturním a kreativním průmyslu, kde mohou zpřístupněné objekty sloužit jako inspirace pro marketing, vzdělávání, gamifikaci nebo umělou inteligenci.

http://doi.org/10.52036/1335793X.2024.2.22-31

ÚVOD

České digitální knihovny uchovávají obrovské množství informací. Ačkoliv většina je ve formě textu, významná část kulturního dědictví je zachycena také graficky – v podobě fotografií, kreseb, schémat, map a dalších grafických prvků. Postupná digitalizace knihovních fondů spojená s optickým rozpoznávání znaků (OCR) umožňuje snadné fulltextové vyhledávání a otevírá tak veřejnosti skryté textové kulturní dědictví. Projekt „Orbis Pictus – oživení knihy pro kulturní a kreativní odvětví" si klade za cíl podobným způsobem zpřístupnit veřejnosti grafický, přesněji netextový obsah českých digitálních knihoven. Projekt byl schválen Ministerstvem kultury ČR v rámci programu NAKI III pod ID DH23P03OVV033 a jeho řešení probíhá v letech 2023–2027 za spolupráce Knihovny AV ČR, v. v. i., Moravské zemské knihovny v Brně, Národní knihovny České republiky a Vysokého učení technického v Brně.

Netextové objekty – všechny grafické elementy, ale také vzorce, rovnice, tabulky a další entity, které nelze snadno rozpoznat pomocí běžných procesů OCR – bude možné s využitím metod strojového učení identifikovat v digitalizovaných dokumentech, kategorizovat je dle typu a doplnit o kontextové informace, které usnadní jejich vyhledávání a umožní rozšíření nabídky služeb digitálních knihoven.

V prvních dvou letech projektu vznikala metodika zpracování obrazových dokumentů, která vymezuje způsoby a rozsah kategorizace netextových objektů v digitalizovaných dokumentech a možnosti jejich dalšího využití. Paralelně s metodikou je vyvíjen systém pro automatickou anotaci objektů na digitalizovaných stranách nazvaný AnnoPage. Ten bude na bázi strojového učení automaticky analyzovat obsah jednotlivých digitalizovaných stran, identifikuje na nich oblasti obsahující netextové objekty, určí jejich kategorii a doplní ke každému objektu stručný textový popis. Rozpozná-li AnnoPage v netextovém objektu osobu, bude připraven použít další vyvíjený nástroj – PeopleGator. Ten na něj naváže a opět pomocí metod strojového učení umožní indexaci a vyhledání osob podle obličeje. Z osob identifikovaných pomocí AnnoPage a PeopleGatoru vznikne posléze databáze osob identifikovaných v digitálních knihovnách, která bude propojovat digitální knihovny, národní a mezinárodní databázi autorit a další otevřené zdroje informací. Nástroje vyvinuté v rámci projektu budou ve finální fázi implementovány v rozhraní České digitální knihovny[1], která slouží jako agregátor digitálních knihoven provozovaných v systému Kramerius na území ČR.

Cílem prvních dvou let projektu bylo vytvořit vstupní analýzy současného stavu poznání a zejména změn, které nastaly od jeho přípravy. Důraz byl kladen také na podrobný rozbor datových základen digitálních knihoven institucí zapojených do projektu. Současně byly rozpracovány návrhy komunikačních rozhraní mezi nově vyvíjenými aplikacemi a již zavedenými systémy digitálních knihoven, včetně způsobů efektivní výměny dat mezi nimi. Klíčovým výstupem roku 2024 potom byla metodika pro kategorizaci netextových částí dokumentů, ve které byl navržen postup a rozsah kategorizace netextových objektů a způsob zápisu informací o těchto objektech do strukturovaných metadat.

TŘI KROKY K EFEKTIVNÍMU VYHLEDÁVÁNÍ NETEXTOVÝCH OBJEKTŮ V DIGITALIZOVANÝCH DOKUMENTECH

Prvním krokem umožňující vyhledávání netextových objektů v digitalizovaných dokumentech je jejich jednoznačná identifikace a lokalizace. Následně je možné z detekovaných objektů extrahovat jejich popis a poté získané informace za účelem vyhledávání indexovat. Proces rozpoznávání netextových objektů tak lze při určitém zjednodušení rozdělit na tři hlavní úkoly: detekovat objekt, extrahovat jeho popis a zajistit jeho vyhledatelnost.

METODY DETEKCE

Rozpoznávání grafických prvků v digitalizovaných předlohách je součástí procesů pro rozpoznání tištěného (OCR, Optical Character Recognition) nebo ručně psaného textu (HTR, Handwritten Text Recognition). Tyto procesy běžně vymezují části stran, které obsahují text k rozpoznání, a oddělují jej od částí, které text neobsahují. Pokročilejší analýza objektů na stránce se proto často objevuje v projektech zaměřených na vylepšení rozpoznávání textů.[2]

Moderní strojové učení nabízí různé přístupy k detekci objektů v obrazech. V zásadě všechny ovšem používají konvoluční neuronové sítě k analýze vstupního obrazu a algoritmy, které na základě výstupů vytvářejí výsledné detekce. Konečným produktem těchto procesů jsou osově zarovnané obdélníky s informací o kategorii detekovaného objektu a případně o pravděpodobnosti označující míru nejistoty detekce. Mezi nejčastěji využívané metody patří například YOLO (Reis et al. 2023), SSD (Liu et al. 2016) nebo Faster RCNN (Ren et al. 2015). V projektu Orbis Pictus jsme se rozhodli pro metodu YOLO.

Aby byla detekce úspěšná, je třeba modely strojového učení trénovat na vhodné a dostatečně velké datové sadě. Jedním z prvních projektů, který se zaměřil na grafický obsah digitalizovaných tiskových materiálů

a jejich automatickou anotaci, byl Newspaper Navigator (Lee et al. 2020). Ten stál na výstupech projektu Beyond Words, ve kterém vznikaly popisy pro netextové části digitalizátů prostřednictvím crowdsourcingu. Dataset z tohoto projektu obsahuje zejména popisky získané z textového okolí obrázků. Existuje také několik veřejně dostupných datových sad s anotacemi obrázků z historických dokumentů, jako je IlluHisDoc (Monnier a Aubry 2020), HORAE (Boillet et al. 2019) nebo HBA (Mehri et al. 2017). Obrazová data v těchto sadách ale nejsou dělena do dostatečně podrobných kategorií a některé typy netextových objektů nejsou anotovány vůbec. Mezi obecnější datové sady potom patří například ImageNet s 14,1 miliony obrázků (Stanford Vision Lab 2020), COCO – Common Objects in Context s 200 miliony obrázků s popisky (Lin et al. 2014) nebo COYO-700M (Byeon et al. 2022) se 747 miliony popsaných objektů.

Problematikou identifikace obrázků v digitalizovaných publikacích zkoumali pracovníci MZK v roce 2021[3], když porovnávali výstupy existujícího OCR (ABBYY Recognition Server v různých verzích) s nástroji strojového učení (například zmíněný Newspaper Navigator nebo dhSegment). Závěry jejich testů ukázaly, že se stávajícími nástroji nelze při zpracování obrazových dat z digitálních knihoven dosáhnout dostatečné spolehlivosti a jejich praktické nasazení bez dalšího vývoje tak není možné. V rámci projektu Orbis Pictus jsme se i proto rozhodli vytvořit vlastní ručně anotovanou datovou sadu, která zahrnuje dostatečný počet kategorií pro netextové objekty vyskytující se v digitalizovaných dokumentech s cílem zlepšit jejich vyhledatelnost a použitelnost.

KATEGORIE OBJEKTŮ

Detekované netextové objekty jsme v rámci v projektu Orbis Pictus rozdělili do 25 kategorií. Tyto kategorie byly vytvořeny na základě kritérií zohledňujících zejména možnosti využití pro kreativní průmysl, využitelnost pro další výzkum a celkové zlepšení kvality zpřístupnění obsahu digitálních knihoven. Výběr kategorií není konečný – v budoucnu lze přidat nové kategorie, vyplyne-li to z potřeb uživatelské komunity.

Aby bylo zajištěno kvalitnější zpřístupňování obsahu digitálních knihoven a zabránilo se záměně s jinými grafickými prvky, jsou mezi kategorie zařazeny také takové, které přesahují definici netextových objektů. Jedná se například o chemické vzorce a rovnice, matematické výrazy a rovnice či tabulky. Přestože mají tyto prvky spíše textový charakter, smysl jejich detekce spočívá v tom, že mohou být vyloučeny z výsledků vyhledávání grafických objektů v uživatelském rozhraní. Jejich identifikace je ale také důležitá pro tvorbu datasetů, které lze využít ke zlepšení rozpoznávání jejich obsahu a následné extrakci dat pro další použití.

V rámci definic kategorií a jejich anotování předpokládáme, že všechny detekované objekty mohou obsahovat zanořené objekty z jiných kategorií, které jsou detekovány samostatně. Například razítko může obsahovat samostatně detekovaný heraldický symbol.

Pro každou ze stanovených kategorií bylo následně nutné zajistit dostatek příkladů – anotací na stranách digitalizovaných dokumentů. K anotování jsme využili open-source nástroj Label Studio[4], který nabízí mimo jiné šablony pro detekci objektů a jejich klasifikaci. Při nahrávání obrazových dat do tohoto nástroje je možné předem nastavit predikce, tedy vyznačit, co by mělo být na daném obrazu nalezeno. Tento přístup urychluje práci, neboť jednotlivé anotace stačí pouze zkontrolovat a není nutné je vytvářet zcela ručně. Tento postup byl například využit při anotování kategorií typických pro staré tisky, kde knihovědci předem opatřili data popisem specifikujícím, jaký typ knižní ilustrace nebo dekoru se na stránkách nachází (rozlišujeme například viněty, vlysy či signety). Po vytvoření nebo úpravě všech anotací je možné vytvořená data jednoduše vyexportovat z Label Studia v několika standardizovaných formátech, včetně formátu vhodného pro trénování námi zvoleného YOLO detektoru.

V první fázi jsme do Label Studia nahráli stovky stránek z veřejně dostupných dokumentů z Národní digitální knihovny ČR, Digitální knihovny Moravské zemské knihovny v Brně a Digitální knihovny AV ČR. Na těchto stranách ručně anotovalo 16 pracovníků jednotlivé objekty. Na základě vytvořených anotací byl natrénován první model detektoru YOLO, který byl použit k vytvoření automatických detekcí na další sadě stránek. Tyto detekce byly ručně zkontrolovány a případně upraveny. Pracovníkům byly předkládány ty strany, na kterých byla míra jistoty detekce nižší než optimální hodnota nastavená v detektoru. Tímto způsobem se upřednostňovaly takové typy stran, které nebyly dostatečně zastoupeny v původní sadě a kde model nefungoval optimálně, zatímco typy netextových objektů, které byly rozpoznávány spolehlivě, se už pracovníkům nezobrazovaly. Anotační proces lze opakovat iterativně, dokud detektor nedosáhne požadované úrovně přesnosti.

EXTRAKCE METADAT

Identifikace netextových objektů na digitalizovaných stranách je ovšem pouze prvním krokem. Dále je nutné získat informace o jejich obsahu a vytvořit optimální

metadatový záznam, který umožní jejich efektivní vyhledávání. V projektu Orbis Pictus popisujeme objekty pomocí vektorových reprezentací a textových popisů.

Vektorová reprezentace (sémantický embedding) není přímo interpretovatelná člověkem, protože se jedná o sadu reálných čísel, jejichž jednotlivé prvky samy o sobě nenesou konkrétní význam. Nicméně tyto vektory lze použít k výpočtu podobnosti mezi dvěma objekty, a to jak mezi dvěma obrázky, tak mezi obrázkem a textem.

Pro tvorbu vektorových reprezentací lze využít existující modely, jako jsou CLIP (Radford et al. 2021) nebo ALIGN (Jia et al. 2021). Tyto modely používají neuronové sítě k vytváření vektorů jak pro obrázky, tak pro texty. Oba modely jsou trénovány na velkých datových sadách obrázků a jejich textových popisů, aby se vektory obrázků a jejich popisů co nejvíce shodovaly.

Doplnění vektorů o textový popis (v našem případě klíčová slova a krátký popis volným textem) zjednodušuje vyhledatelnost, protože textový popis může sloužit jako alternativní reprezentace k vektorovým datům. Textové popisy obrázků lze získat buď generováním pomocí natrénovaného modelu, nebo hledáním v okolním textu.

Extrahovat textový popis umožňují generativní neuronové sítě podobné současným velkým jazykovým modelům. Text je generován na základě obsahu samotného netextového objektu. Existují datové sady, na kterých lze trénovat modely umělé inteligence, a také veřejně dostupné natrénované modely.[5] Problém těchto sad spočívá v tom, že popisy obrázků jsou v angličtině a obsahují především současné obrázky a fotografie. Výsledky modelů trénovaných na těchto datech proto nemusí být dostatečně přesné pro obrázky z historických dokumentů. První problém lze řešit pomocí překladače, druhý si může vynutit dotrénováním modelu na vlastní datové sadě historických obrázků a jejich popisů.

Pro získání popisu objektu na základě textu, který tento objekt obklopuje, lze použít podobný přístup jako při vyhledávání obrázků – na základě vektorových reprezentací lze vyhledávat relevantní textové řádky či věty. K tomuto procesu může přispět také informace o pozici obrázku na stránce, neboť popisy se často nacházejí v jeho blízkosti. Alternativou je detektor, který se naučí identifikovat místa na stránce, kde se nacházejí popisy obrázků, podobně jako detektor pro netextové objekty.

Je důležité mít na paměti, že neuronové sítě pro extrakci vektorových reprezentací i textových popisů jsou převážně trénovány na obsahu z internetu. Jejich schopnosti a předpojatosti odpovídají tomuto kontextu, a proto nejsou schopny zachytit informace, které se na internetu vyskytují jen zřídka. Tyto modely tak například nepoznají méně známé historické osobnosti, místa nebo specifické objekty.

Současné metody umožňují trénovat modely, které poskytují společnou sémantickou reprezentaci (embedding) pro obrázky i texty. Například již zmíněný CLIP, se kterým nyní pracujeme v rámci projektu, je trénován na velkých datových sadách obrázků a jejich textových popisků sesbíraných z internetu. Díky tomu je schopen rozumět sémantickému obsahu obrázků v rozsahu blížícímu se lidskému chápání (Geirhos et al. 2021). Tento model tak lze snadno použít ke generování textových popisů obrázků pro standardní fulltextové vyhledávání.

ZÁPIS METADAT

Textové popisy a vektory použitelné pro vyhledávání netextových objektů musí být zapsány ve strukturovaném a ideálně interoperabilním formátu, ze kterého je možné je efektivně indexovat pro potřeby digitálních knihoven. V projektu Orbis Pictus jsme zvolili formát ALTO (Analyzed Layout and Text Object), který v současnosti spravuje Kongresová knihovna. Účelem formátu ALTO je vytvořit mapu vztahů pro každou stranu digitalizovaného dokumentu (Kočišová et al. 2023). ALTO umožňuje přesnou reprezentaci struktury stránek včetně umístění textu, grafických objektů a dalších prvků. Díky jeho využití je například již nyní možné uživatelům digitálních knihoven při prohlížení digitalizovaných dokumentů zvýraznit konkrétní oblasti na stránce jako hledaná slova, celé bloky textu nebo ilustrace či jiné grafické objekty. ALTO také umožňuje stanovení pořadí čtení jednotlivých segmentů na stránce, což je užitečné zejména v případech, kdy text není uspořádán lineárně nebo kdy stránka obsahuje více sloupců či oddělených částí textu (Lehečka 2023).

Formát ALTO je dnes nejen v ČR běžně využíván v procesech digitalizace, kdy je čistý text rozpoznaný technologií OCR převeden do ALTO XML a k rozpoznaným textovým blokům či grafickým objektům jsou připojeny souřadnice určující, kde přesně se na stránce grafický objekt, ilustrace či tabulka vyskytuje (Kočišová et al. 2023). Tato funkce je v rámci projektu Orbis Pictus klíčová pro správné zaznamenání netextových objektů a jejich umístění.

Významným důvodem pro volbu formátu ALTO je také skutečnost, že je od roku 2012 součástí Standardů Národní digitální knihovny[6], které slouží jako jednotný formát pro digitalizační projekty v České republice

a jako takové ho využívají české digitální knihovny pracující se systémem Kramerius. Další významnou výhodou formátu ALTO je možnost vnořit do jeho struktury data v jiném XML formátu. To umožňuje použít pro detailnější popis obrázků, ilustrací či jiných grafických objektů externí schémata. Tato flexibilita je pro projekt Orbis Pictus zásadní, neboť pomocí jediného XML souboru tak lze popsat nejen text a jeho umístění na stránce, ale také například autorství obrázků nebo identifikaci osob zobrazených na ilustracích.

VYHLEDÁVÁNÍ

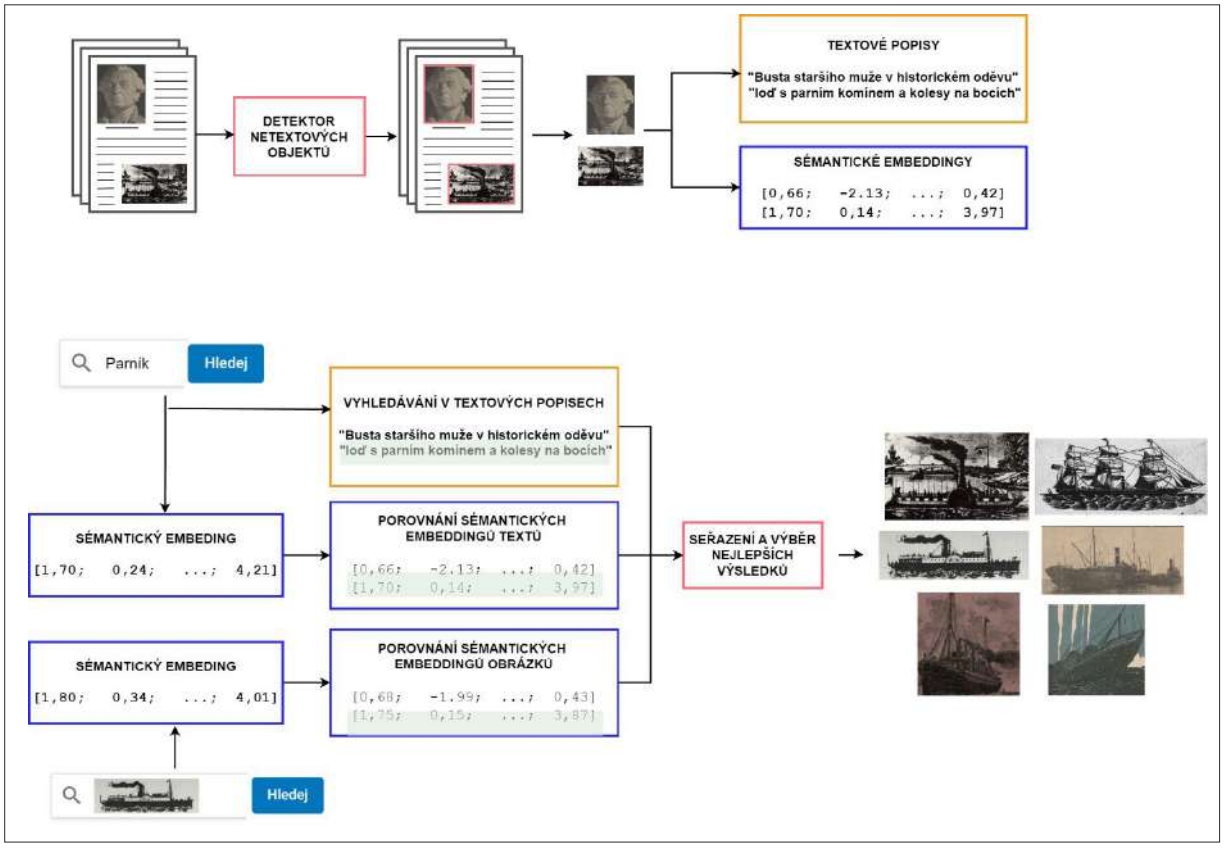
Identifikované netextové objekty s metadatovým popisem je možné vyhledávat textovým dotazem i obrázkem. Textový dotaz se využívá pro klasické fulltextové vyhledávání ve vygenerovaných textových popisech objektů. Pro většinu kategorií grafických objektů (kromě tabulek, matematických a chemických formulí apod.) lze ale také použít tzv. vektorové vyhledávání. Výsledky obou způsobů vyhledání je potom možné spojit podle míry podobnosti (jedno reálné číslo), které produkuje vektorové vyhledání a skóre full-textového vyhledání.

Vektorové vyhledávání funguje tak, že každý identifikovaný objekt je nejprve pomocí neuronové sítě převeden na vektorovou reprezentaci – například na vektor sestávající z 768 desetinných čísel, který jej charakterizuje. Každý vyhledávací dotaz uživatele je poté stejnou neuronovou sítí také převeden na vektorovou reprezentaci o stejné velikosti a v databázi jsou pak nalezeny vektory objektů nejpodobnější k vektoru dotazu pomocí euklidovské nebo kosinové podobnosti. Celý tento proces znázorňuje obrázek 1.

DOSAVADNÍ VÝSLEDKY

V prvních dvou letech projektu Orbis Pictus byly provedeny vstupní analýzy mapující současný stav poznání od doby přípravy projektu a analýza datové základny. Byly rozpracovány představy o komunikačních rozhraních mezi jednotlivými aplikacemi, které jsou výstupem projektu, i jejich napojení na fungující digitální knihovny. Především byla dokončena certifikovaná metodika pro zpracování obrazových dokumentů[7] standardizující práci s netextovými entitami, která stanovila jejich kategorizaci, způsob jejich identifikace a způsob zápisu získaných informací do strukturovaných metadatových záznamů za využití schématu ALTO.

Metodika je určena zejména pro tři skupiny uživatelů: kurátory digitálních knihoven, kteří jsou odpovědní za sprá-



Obr. 1 Schéma systému pro vyhledávání netextových objektů v dokumentech

vu a zachování digitálních dat, badatele v oblasti humanitních věd, kteří využívají digitalizovaná data pro své výzkumy, a další potenciální uživatele, jako jsou zástupci kreativního průmyslu, kteří v metodice najdou informace o tom, jak využít datovou základnu knihoven pro své projekty včetně zhodnocení obrazové kvality obsahu digitálních knihoven a právní aspekty spojené s komerčním využitím digitálních dat.

Metodika vysvětluje datovou základnu českých digitálních knihoven, včetně historie jejich budování a využívaných standardů, což mělo a stále má dopad na kvalitu obrazů i metadata, která jsou skrze české digitální knihovny dostupná. Nejrozsáhlejší část metodiky se věnuje kategorizaci netextových objektů v rámci projektu. Další části popisují způsoby jejich detekce, popisu a vyhledávání, doporučené standardy metadat (zejména ALTO XML) a jejich využití. Součástí metodiky je i popis možností využití výsledků projektu pro digitální humanitní vědy a kreativní průmysl.

V rámci projektu Orbis Pictus bylo definováno 25 kategorií, do nichž jsou jednotlivé netextové objekty řazeny. Zmíněná metodika poskytuje jednotný návod a terminologii pro kategorizaci a popis netextových objektů použitou v rámci vývoje nástrojů pro projekt. Tato kategorizace si ovšem klade také za cíl usnadnit odborný i veřejný výzkum obsahu digitálních knihoven tím, že uživatelům pomůže orientovat se v jednotlivých definovaných kategoriích objektů.

Pro každou kategorii, jejichž výčet přináší tabulka 1, byla v metodice vedle obecně platné definice daného objektu stanovena specifická definice pro účely projektu, která upřesňuje, jak konkrétně byla nastavena pravidla pro jeho zařazení do kategorie. Pro každou ka-

SOUHRNNÁ KATEGORIE	DETEKOVANÁ KATEGORIE
	Čárové a QR kódy
	Exlibris
	Fotografie
	Grafy
	Karikatura a komiks
Knižní dekor	Viněty Vlysy Iniciály Signet Ostatní knižní dekor
	Mapy
	Notové zápisy
	Ozdobné nápisy
	Razítka
	Reklamy
	Rukopisné vpisky
	Symboly, loga a erby
	Tabulky
Výkresy	Půdorysy Schémata Geometrické výkresy Ostatní technické výkresy
Výrazy, vzorce a rovnice	Matematické výrazy a rovnice Chemické vzorce a rovnice
	Obrázek

Tab. 1 Kategorie netextových objektů

tegorii jsou v metodice popsány také důvody, proč je její existence relevantní z hlediska dalšího využití a výzkumu a doplněny jsou i konkrétní příklady, včetně příkladů mezních. Obsahově blízké kategorie jsou potom řazeny pod souhrnné názvy, které však samy o sobě nejsou detekovanými kategoriemi. Do souhrnné kategorie Obrázek pak byla zařazena jakákoliv netextová vizuální entita, kterou nebylo možné přidat do žádné z ostatních stanovených kategorií.

Pomocí výše popsaného iterativního procesu anotování digitalizovaných stran byla vytvořena datová sada pracující se všemi 25 kategoriemi nazvaná AnnoPage Dataset. Dataset obsahuje volně dostupné stránky z digitálních knihoven Národní Knihovny ČR, Moravské zemské knihovny v Brně a Knihovny AV ČR a také stránky z již existujících volně dostupných datových sad.[8] Anotace jsou ve formátu, který se dá přímo použít pro trénování YOLO detektoru. Celkem datová sada obsahuje 7550 stran (z toho 5690 z českých digitálních knihoven) s celkem 27 904 anotacemi. Obrázek 2 znázorňuje histogram rozložení četnosti jednotlivých kategorií. Datová sada je veřejně dostupná pod licencí CC BY 4.0 a ke stažení z repozitáře Zenodo.[9]

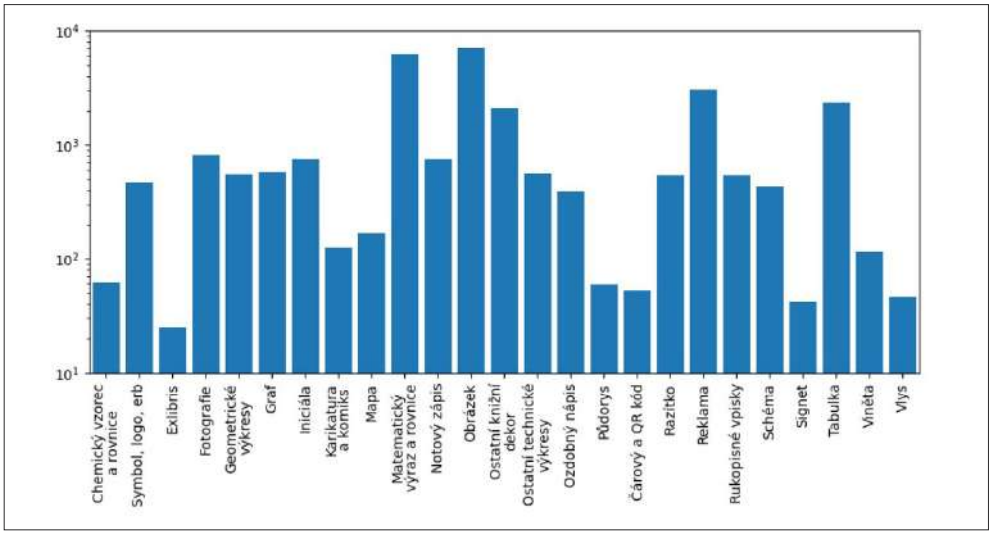
Součástí metodiky pro zpracování obrazových dokumentů je podrobný předpis pro metadatový zápis grafických objektů v ALTO XML. Ten předpokládá použití formátu ALTO XML ve verzi 4.4. Toto schéma umožňuje k popisu objektů využít kořenový element <Tags> a jeho dceřiné elementy, které mohou nést podrobnější informace o rozpoznaných objektech:

- Pro obrázek se využívá element <LayoutTag>.
- Pro nadpis obrázku <StructureTag>.
- Pro autora či vydavatele <RoleTag>.
- K pojmenování osob či institucí <NamedEntityTag>.
- Pro pasáž textu, která s obrázkem souvisí nebo na něj odkazuje, je možné využít <OtherTag>.

Každý z uvedených subelementů elementu <Tags> může skrze element <XmlData> obsahovat podrobnější bibliografický popis objektu využívající libovolné XML schéma. Pro projekt Orbis Pictus bylo zvoleno schéma MODS (Metadata Object Description Schema), které je široce využívané mezinárodní komunitou a je taktéž součástí Standardu NDK. MODS vychází z katalogizačního standardu MARC 21, ale oproti němu je lidsky lépe čitelný (Kočišová et al. 2023). V současnosti je aktuální verze MODS 3.8 ze září 2022.

Zatímco samotné schéma ALTO popisuje zejména pozici objektu na straně, právě v rámci vnořeného schématu MODS jsou zapsány bližší údaje o kategorii grafického objektu (konkrétně v elementu <genre>), klíčová slova (<subject>/<topic>) a shrnutí obsahu obrázku (element <abstract>). Vektor určený pro vyhledávání objektů součástí ALTO není, vzhledem ke své velikosti. Jeho distribuce pro další práci s ním v rámci digitálních knihoven se předpokládá pomocí jednoduchého zápisu v souboru formátu JSON.

Výše uvedené části – jasně definované kategorie netextových objektů, datová sada umožňující spolehlivé fungování jejich detekce a standardizovaný způsob zápisu informací o detekovaných objektech do interoperabilního schématu – konstituují dohromady základy nástroje AnnoPage, který je jedním ze dvou inovativních technických nástrojů vyvíjených v rámci projektu Orbis Pictus.



Obr. 2 Histogram rozložení počtu jednotlivých kategorií ve vytvořené datové sadě

AnnoPage bude identifikovat netextové části stránek a přiřadí jim automaticky textový popis na základě jejich obsahu a souvisejícího textu. Výstupy nástroje budou vhodné k indexaci a použití při vyhledávání přímo v digitálních knihovnách. Nástroj bude dokončen v průběhu roku 2025.

V rámci pokračování projektu Orbis Pictus naváže na nástroj AnnoPage nástroj PeopleGator, jehož vstupem budou ty objekty, ve kterých AnnoPage vyhodnotí, že zobrazující osoby. PeopleGator, který bude dokončen v roce 2026, využije metody strojového učení k identifikaci těchto osob a pro osoby vyskytují se na více vyobrazeních umožní vytvoření virtuálního grafu propojujícího dokumenty zobrazující tyto osoby. Tato funkcionalita bude využita jak přímo při vyhledávání podle obrazové předlohy, tak pro vytvoření databáze identifikovaných osob a jejich výskytů v dokumentech, což umožní vyhledávání obličejů osob podle jména.

Databáze osob identifikovaných v digitálních knihovnách bude obsahovat jednak vyobrazení osob nalezených v digitálních knihovnách pomocí AnnoPage a PeopleGatoru, ale pracovat bude také s externě získanými vyobrazeními osob již identifikovaných z různých otevřených zdrojů (Wikimedia, Obalkyknih.cz a dalších). Díky tomuto propojení bude možné poskytnout uživateli široký výběr různých vyobrazení jím hledané osoby, ale také jej propojit s dokumenty, které dané vyobrazení obsahují. Databáze je zamýšlena jako backend systém poskytující data prostřednictvím API klientské aplikaci, přičemž API bude sloužit také k plnění databáze. Díky jeho existenci bude navíc velmi usnadněna integrace databáze do dalších nástrojů.

Databáze identifikovaných osob bude dokončena v roce 2026 spolu se softwarovým řešením „Kramerius Orbis Pictus“. To bude představovat novou online aplikaci integrující nástroje vyvinuté v rámci projektu a jejich výstupy, umožňující nadstandardní využití netextových částí dokumentů uložených v digitálních knihovnách provozovaných v systému Kramerius. Aplikace bude komunikovat se systémem Kramerius i dalšími nástroji prostřednictvím API, což zajistí její nezávislý provoz. Její součástí budou také praktické nástroje pro pokročilou práci s grafickým obsahem digitální knihovny a jeho dalším využitím včetně možností pokročilé práce s výřezy dokumentů a jejich snadné sdílení, vyhledávání v kategoriích grafických entit identifikovaných v dokumentech či virtuální propojení dokumentů zobrazujících stejné osoby.

Všechny výše uvedené části budou integrovány do České digitální knihovny, což umožní jejich využívání všemi zúčastněnými knihovnami prostřednictvím rozhraní „Česká digitální knihovna – Orbis Pictus“. Nástroje tak budou dostupné všem knihovnám do České digitální knihovny zapojeným. Tím se významně rozšíří také množství obsahu nabízeného díky projektu Orbis Pictus. Poloprovoz tohoto systému je naplánován na konec projektu v roce 2027.

ZÁVĚR

Projekt „Orbis Pictus – oživení knihy pro kulturní a kreativní odvětví“ se zaměřuje na zpřístupnění grafického, konkrétně netextového obsahu českých digitálních knihoven veřejnosti. Cílem je pomocí metod strojového učení identifikovat, kategorizovat a popsat netextové objekty, jako jsou fotografie, kresby, schémata či vzorce, které OCR technologie nerozpozná.

Projekt v prvních dvou letech svého řešení úspěšně splnil dosavadní cíle, které vedle vstupních analýz zahrnovaly především vypracování metodiky pro kategorizaci netextových částí dokumentů. Tato metodika je klíčovým základem pro nástroj AnnoPage, jenž bude sloužit k automatické anotaci netextových objektů v digitalizovaných dokumentech. Metodika stanovila rozsah a popis kategorií těchto objektů, umožnila vytvoření a zveřejnění datové sady pro trénování detektorů využívajících strojové učení a detailně popsala způsob zápisu metadat, který bude pro popis objektů používán.

Shromáždění a kategorizace velkého množství nejen grafických dat a jejich snadné zpřístupnění činí Orbis Pictus unikátním projektem na poli digital humanities. Uplatnění najdou výsledky jak přímo ve využití identifikovaných grafických objektů a jejich metadat, tak celých datových sad pro další výzkum. Orbis Pictus představuje také unikátní příležitost v oblasti kulturního a kreativního odvětví. Kreativních průmysly představující odvětví ekonomiky založené na individuální či skupinové kreativitě, dovednostech a talentu. Grafické objekty zpřístupněné skrze projekt mohou sloužit jako zdroj inspirace nebo přímo využitelné podklady pro různá odvětví jako například marketing, vzdělávání, gamifikace, interaktivní zážitky či široké pole umělé inteligence a generativních modelů.

POUŽITÁ LITERATURA

BOILLET, Mélodie; BONHOMME, Marie-Laurence; STUTZMANN, Dominique a Christopher KERMORVANT, (2019). HORAE: *an annotated dataset of books of hours. In: HIP '19: Proceedings of the 5th International Workshop on Historical Document Imaging*

and Processing. Online. Spojené státy americké: Association for Computing Machinery, s. 7–12. ISBN 1-4503-7668-1. Dostupné z: https://doi.org/10.1145/3352631.3352633. [cit. 2024-08-29].

BYEON, Minwoo; PARK, Beomhee; KIM, Haecheon; LEE, Sungjun; BAEK, Woonhyuk et al., (2022). *Coyo-700m: Image-text pair dataset*. Online repozitář. Dostupné z: https://github.com/kakaobrain/coyo-dataset. [cit. 2024-09-24].

GEIRHOS, Robert; KANTHARUJU, Narayanappa; MITZKUS, Benjamin THIERINGER, Tizian; BETHGE, Matthias WICHMANN, Felix et al., (2021). Partial success in closing the gap between human and machine vision. In: *Advances in Neural Information Processing Systems*. Online. 34: 23885–23899. Dostupné z: https://proceedings.neurips.cc/paper/2021/file/c8877cff22082a16395a57e97232bb6f-Paper.pdf. [cit. 2024-09-24].

JIA, Chao; YANG, Yinfei; XIA, Ye; CHEN, Yi-Ting; PAREKH, Zarana et al., (2021). Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In: ZHANG, Tong, MEILA, Marina (ed.). *Proceedings of the 38th International Conference on Machine Learning*. Online. Spojené státy americké: PMLR, s. 4904–4916. ISBN 9781713845065. ISSN 2640-3498. Dostupné z: https://doi.org/10.48550/arXiv.2102.05918. [cit. 2024-09-24].

KOČIŠOVÁ, Pavlína; VAŠEK, Zdeněk; JIROUŠEK, Václav; KOPSKÝ, Vojtěch; BILWACHS, Jan et al., (2023). *Zachováno navěky?: teorie a praxe dlouhodobého uchování digitálních dokumentů*. Praha: Národní knihovna České republiky. ISBN 978-80-7050-791-9. Dostupné také z: https://www.nkp.cz/soubory/ostatni/zachovano_naveky.pdf. [cit. 2024-09-24].

LEE, Benjamin Charles Germain; MEARS, Jaime; JAKEWAY, Eileen; FERRITER, Megha; ADAMS, Chris Adams, YARASAVAGE, Nathan et al., (2020). *The Newspaper Navigator Dataset: Extracting And Analyzing Visual Content from 16 Million Historic Newspaper Pages in Chronicling America*. Online. ArXiv preprint ArXiv:2005.01583. Dostupné z: https://doi.org/10.48550/arXiv.2005.01583. [cit. 2024-29-24].

LEHEČKA, Boris, (2023). *Blýskání na lepší data z českých digitálních knihoven*. Časopis pro moderní filologii. Online. roč. 105, č. 2, s. 274–291. ISSN 2336-6591. Dostupné z: https://doi.org/10.14712/23366591.2023.2.7. [cit. 2024-09-24].

LIN, Tsung-Yi; MAIRE, Michael; BELONGIE, Serge; HAYS James; PERONA, Pietro; RAMANAN, Deva et al., (2014). Microsoft coco: Common objects in context. In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6-12.9., 2014, Proceedings, Part V 13*. Springer International Publishing, s. 740–755. Dostupné z: https://link.springer.com/chapter/10.1007/978-3-319-10602-1_48. [cit. 2024-09-24].

LIU, Wei; ANGUELOV, Dragomir; ERHAN, Dumitru; SZEGEDY, Christian; REED, Scott et al., (2016). SSD: Single Shot MultiBox Detector. Online. In: LEIBE, Bastian; MATAS, Jiri; SEBE, Nicu a Max WELLING, (ed.). *Computer Vision – ECCV 2016. Lecture Notes in Computer Science*. Cham: Springer International Publishing, s. 21–37. ISBN 978-3-319-46447-3. Dostupné z: https://doi.org/10.1007/978-3-319-46448-0_2. [cit. 2024-09-24].

MEHRI, Maroua; HÉROUX, Pierre; MULLOT, Rémy; MOREUX, Jean-Philippe; COÜASNON, Bertrand et al., (2017). HBA 1.0: A Pixel-based Annotated Dataset for Historical Book Analysis. In: *HIP '17: Proceedings of the 4th International Workshop on Historical Document Imaging and Processing*. Online. Spojené státy americké: Association for Computing Machinery, s. 107–112. ISBN 978-1-4503-5390-8. Dostupné z: https://doi.org/10.1145/3151509.3151528. [cit. 2024-08-29].

MONNIER, Tom a Mathieu AUBRY, (2020). DocExtractor: An off-the-shelf historical document element extraction. In: *2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. Online. Německo: ICFHR, s. 91–96. ISBN 978-1-7281-9966-5. Dostupné z: https://doi.org/10.1109/ICFHR2020.2020.00027. [cit. 2024-09-24].

ARES OLIVEIRA, Sofia; SEGUIN, Benoit a KAPLAN, Frederic, (2018). DhSegment: A Generic Deep-Learning Approach for Document Segmentation. In: *16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. Online. s. 7–12. ISBN 978-1-5386-5875-8. Dostupné z: https://doi.org/10.1109/ICFHR-2018.2018.00011. [cit. 2024-09-24].

RADFORD, Alec; WOOK KIM, Jong; HALLACY, Chris; RAMESH, Aditya; GOH, Gabriel et al., (2021). Learning transferable visual models from natural language supervision. In: MEILA, Marina a Tong ZHANG, (ed.). *Proceedings of the 38th International Conference on Machine Learning*. Online. Spojené státy americké: PMLR, s. 8748–8763. ISBN 9781713845065. ISSN 2640-3498. Dostupné z: https://doi.org/10.48550/arXiv.2103.00020. [cit. 2024-08-29].

REIS, Dillon; KUPEC, Jordan; HONG, Jacqueline a Ahmad DAOUDI, (2023). *Real-Time Flying Object Detection with YOLOv8. Online*. ArXiv preprint arXiv:2305.09972. s. 1–10. Dostupné z: https://doi.org/https://doi.org/10.48550/arXiv.2305.09972. [cit. 2024-09-24].

REN, Shaoqing; HE, Kaiming; GIRSHICK, Ross a SUN, Jian, (2015). Faster R-CNN: Towards Real-Time Object Detection with

Region Proposal Networks. In: *Advances in Neural Information Processing System*. Online. New York: Curran Associates, s. 1–10. ISBN 9781510825024. ISSN 1049-5258. Dostupné z: https://proceedings.neurips.cc/paper_files/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf. [cit. 2024-09-24].

REZANEZHAD, Vahid; BAIERER, Konstantin; GERBER, Mike; LABUSCH, Kai a Clemens NEUDECKER, (2023). Document Layout Analysis with Deep Learning and Heuristics. In: *Proceedings of the 7th International Workshop on Historical Document Imaging and Processing*. Online. New York, NY, USA: ACM., s. 73–78. ISBN 9798400708411. Dostupné z: https://doi.org/10.1145/3604951.3605513. [cit. 2024-09-24].

SCHUHMANN, Christoph; BEAUMONT, Romain; VENCU, Richard; GORDON, Cade; WIGHTMAN, Ross et al., (2022). LAION-5B: An open large-scale dataset for training next generation image-text models. In: *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*. Online. Spojené státy americké: NIPS, s. 25278–25294. ISBN 9781713871088. Dostupné z: https://doi.org/10.48550/arXiv.2210.08402. [cit. 2024-09-24].

Stanford Vision Lab, (2020). Online. Princeton University, Stanford University. Dostupné z: https://image-net.org/. [cit. 2024-09-24].

ŽABIČKA, Petr; BOGÁR, Jan a Michal TRAN, (2021). *Obrazová data a strojové učení*. Online přednáška. Konference Archivy, knihovny, muzea v digitálním světě 2021 (22. ročník), 30. listopadu – 1. prosince 2021. Dostupné z: https://www.skipcr.cz/sites/default/files/documents/2022-04/31_Zabicka_2021.pdf. [cit. 2024-09-26].

POZNÁMKY

[1] https://ceskadigitalniknihovna.cz/

[2] Takovým projektem je například dhSegment (Oliveira et al. 2018) rozpoznávající ornamentální prvky ve starých tiscích nebo umístěním fotografie na barevném pozadí. Projekt Eynolah (Rezanezhad et al. 2023) dále rozšířil identifikované objekty na stránce např. o nadpisy, iniciály, marginální text nebo tabulky.

[3] Analýza současných nástrojů a možností probíhala v rámci přípravy na podání projektu Orbis Pictus a její závěry byly prezentovány na konferenci Archivy, knihovny a muzea v digitálním světě v prosinci 2021 (Žabička et al. 2021).

[4] https://labelstud.io/

[5] Mezi existující datové sady patří například LAION-5B (Schuhmann et al. 2022), MS-COCO (Lin et al. 2014) a COYO-700M (Byeon et al. 2022).

[6] https://standardy.ndk.cz/ndk/standardy-digitalizace/

[7] Metodika prochází v současnosti procesem certifikace Ministerstva kultury ČR. Bude zveřejněna prostřednictvím českého Národní úložiště šedé literatury, https://nusl.cz/.

[8] Například PRImA Layout Analysis Dataset, PRImA RDCL 2019, PRImA Europeana Newspapers, IlluHisDoc, TexBiG a READ ICDAR2019 cBAD Dataset

[9] https://doi.org/10.5281/zenodo.12788420

■ **Článok bol recenzovaný.**