

Archívy v škatuliach by mohli skončiť

PÍSMENNÉ DEDIČSTVO CHCÚ ZACHRÁNIŤ S POMOCOU UMELEJ INTELIGENCIE

V Štátnom archíve v Košiciach ležia kilometre škatúl s písomnými dokumentmi. Vyhľadávajú ich napríklad občania pri vybavovaní dokumentácie k rôznym životným situáciám alebo historici pri svojom výskume. Projekt ArchAlve spája archív, výskumníka z univerzity a inovačnú spoločnosť, aby otestoval nový spôsob, ako pomocou moderných technológií, nástrojov s podporou umelej inteligencie a ľudskej expertízy zachrániť, roztrieďiť a sprístupniť čo najviac z tohto písomného dedičstva skôr, než bude neskoro. Porozprával o tom MARTIN PEKÁR, profesor slovenských dejín na Filozofickej fakulte Univerzity Pavla Jozefa Šafárika v Košiciach a zodpovedný vedúci projektu.

Archívny dlh sa často spomína ako abstraktné číslo. Čo si pod tým má predstaviť bežný človek?

Najprv musím povedať, že štátne archívy sú dôležitou súčasťou infraštruktúry štátu, len nie sú také viditeľné ako cesty či nemocnice. V archívoch sú uložené dokumenty, ktoré potrebujeme napríklad pri priznaní dôchodku, pri dokazovaní vlastníctva nehnuteľnosti alebo pri overovaní rozhodnutí úradov.

Zároveň sú tam jedinečné historické pramene, ktorých štúdium tvorí základ práce historikov. Archívne dokumenty však majú širší celospoločenský význam, pretože udržiavajú informačnú kontinuitu, a tým prispievajú k chodu štátu a budovaniu našej kolektívnej identity. V neposlednom rade sú súčasťou svetového písomného dedičstva.

Termín archívny dlh znamená, že dokumentov v archíve pribúda rýchlejšie, ako ich archívári stíhajú spracovať a sprístupniť. Ak by sme dnes nespracované škatule s archívnymi dokumentmi postavili do jedného radu, hovoríme v rámci celého Slovenska o desiatkach kilometrov nespracovaných dokumentov. A každý rok pribúdajú ďalšie. Nespracované dokumenty nie sú bežne prístupné verejnosti ani bádateľom. Môžu tiež byť v zlom fyzickom stave, čo urýchľuje ich degradáciu.

Ako konštatujú aj viaceré oficiálne dokumenty, Slovensko v digitalizácii archívov zaostáva. V čom je najväčší problém?

Zjednodušene povedané, problémy sú dva – finančný a technologický. V mnohých krajinách sa digitalizácia archívnych dokumentov robí dlhodobo a systematicky. Na Slovensku zatiaľ digitalizácia prebieha prevažne na báze jednotlivých projektov, pomaly a

nerovnomerne. Často ide len o jednoduché skenovanie bez toho, aby bol digitálny súbor strojovo čitateľný, aby z neho boli vyťažované dáta, aby bol prepojený s archívnymi pomôckami, prípadne ľahko verejne dostupný.

Existencia jednoduchého skenu problém archívneho dlhu nerieši. Ak sa v ňom nedá vyhľadávať, chýbajú metadáta, teda informácie o tom, čo daný dokument obsahuje, k čomu patrí a ako súvisí s inými dokumentmi, vznikajú len „digitálne škatule“.

Archívár aj bádateľ sú na tom v praxi veľmi podobne. Odpadne síce fyzická manipulácia so škatulami a dokumentmi, avšak odborné činnosti ostávajú rovnako náročné. To je jeden z dôvodov, prečo sme sa v projekte ArchAlve rozhodli zamerať na celý proces digitálneho spracovania dokumentov, nie len na rýchlosť skenovania.

Projekt ArchAlve, na ktorom spolupracujete, má tento stav zmeniť. Čo je jeho hlavným cieľom?

Cieľom projektu ArchAlve je vytvoriť a otestovať nový pracovný tok (workflow), ktorý spojí automatizované digitálne spracovanie dokumentov s odbornou prácou archívárov. Nejde o jednorazovú akciu typu „preskenujeme jeden archívny fond“, ale o návrh komplexného riešenia a nového digitálneho štandardu, ktorý sa bude dať použiť opakovane v rôznych archívoch, prípadne aj v inom prostredí.

Naše riešenie má ambíciu pokrývať celý reťazec: od analýzy a výberu dokumentov, cez vysokorýchlostnú digitalizáciu státisícov strán, ich zoradenie, až po rozpoznanie textu, tvorbu a validáciu metadát a návrh elektronickej archívnej pomôcky (inventára). Zároveň berie do úvahy historickú hodnotu dokumentov a archívne štandardy.

Ako taký pracovný tok vyzerá v praxi? V čom je iný než doterajšie postupy?

Na začiatku sa spraví analýza archívnych dokumentov. Zistí sa historická hodnota, typ, formát a fyzický stav dokumentov, ich čitateľnosť či špecifiká písma. Na základe toho sa nastaví plán digitálneho spracovania a technické parametre tak, aby nedošlo k poškodeniu vzácných dokumentov a aby bol celý proces čo najefektívnejší.

Nasleduje fáza vysokorýchlostnej digitalizácie. V projekte chceme výsledne spracovať súbor približne pol milióna strán. Keďže ide o aplikovaný výskum a experimentálny vývoj, budeme testovať rôzne spôsoby, ako zefektívniť celý proces.

V praxi to znamená, že skúsobne naskenujeme až okolo jeden a pol milióna strán. Používame produkčný skener navrhnutý na nepretržitú pre-



Martin Pekár so škatuľou dokumentov, ktoré skener zdigitalizuje za minútu. FOTO: KORZÁR/ELENA DANKOVÁ

vádzku, ktorý pri formáte A4 zvládne za minútu spracovať viac ako dvesto strán. Ide o prístroj a riešenie, vďaka ktorému sa archívári budú môcť viac venovať odbornej práci a menej fyzickej manipulácii s materiálom.

Rozpoznávanie textu zabezpečuje kombinácia OCR a nástrojov umelej inteligencie, ktoré z dokumentov „vyťahujú“ základné údaje, navrhujú metadáta a triedia dokumenty do naprogramovanej štruktúry. Archívár do procesu vstupuje ako expert, ktorý na základe poznania historického kontextu a obsahu dokumentov kontroluje, opravuje a rozhoduje v nejasných prípadoch.

Technológia robí rutinnú, opakovateľnú prácu, človek nesie zodpovednosť za kvalitu a interpretáciu. Výsledkom nebude súbor jednoduchých skenov, ale digitalizáty s metadátami a napokon elektronický inventár, v ktorom sa bude dať systematicky vyhľadávať.

Projekt prepája štátny archív, univerzitu a inovačnú firmu. Prečo je takáto spolupráca dôležitá?

Každý z partnerov prináša do projektu niečo, čo ostatní nemajú. Štátny archív v Košiciach prináša dokumenty, poznanie archívnej praxe a zodpovednosť za písomné dedičstvo. Firma zasa disponuje infraštruktúrou a know-how pri spracovaní veľkých objemov dát.

Moja expertíza ako zodpovedného riešiteľa projektu je prirodzene spojená s Univerzitou Pavla Jozefa Šafárika, kde pôsobím na Filozofickej fakulte ako vedúci Katedry histórie. Pri výskume politických a sociálnych dejín 20. storočia sa venujem otázkam kvality vzdelávania a prenosu digitálnych kompetencií do prostredia humanitných vied. Univerzita však v tomto projekte nefiguruje ako partner, ale ako jeden z odberateľov výsledkov výskumu.

Kombinácia týchto troch

pohľadov umožní vytvoriť riešenie, ktoré je podložené vedecky, realizovateľné technicky a overiteľné v praxi. Nie je preto náhoda, že projekt uspel vo veľmi konkurenčnej grantovej výzve Agentúry pre podporu vedy a výskumu. Som presvedčený, že práve kombinácia verejného, akademického a komerčného sektora bola jednou zo silných stránok projektu. Ide o dobrý príklad prepojenia komerčného sektora s verejným záujmom.

Akú úlohu v projekte zohráva spoločnosť CeDiTra, a. s.?

CeDiTra je žiadateľom projektu a technologicko-inovačným partnerom. Prevádzkuje plnohodnotné digitalizačné centrum s rôznymi typmi skenerov, hardvérovou a softvérovou infraštruktúrou, bezpečným úložiskom dát, ktoré spĺňa požiadavky archívnej legislatívy. Bez tejto nákladnej infraštruktúry by sa výskum a vývoj nemohli realizovať.

Zároveň samotný projekt nie je financovaný iba z verejných zdrojov. CeDiTra ho spolufinancuje. Dáva k dispozícii nielen vlastnú infraštruktúru. Ale v zmysle pravidiel grantovej výzvy znáša i časť nákladov na techniku, vývoj a odborný personál.

Pre firmu je to dlhodobá investícia do výskumu a do riešení, ktoré bude vedieť neskôr ponúknuť ďalším archívom a iným pamäťovým inštitúciám. Pre verejný sektor to znamená, že výstupy projektu vznikajú v inovatívnom prostredí, ktoré je pripravené preniesť ich do praxe.

Pre spoločnosť CeDiTra nejde o prvú spoluprácu so štátnym archívom. Už pred podaním projektu ArchAlve začala na základe zmluvy bezodplatne pilotne digitalizovať fond Košická župa.

Zdôrazňujete, že súčasťou projektu je vzdelávanie. Čo to konkrétne znamená pre archívárov?

Technológie samy osebe nestačia. Ak chceme, aby sa vý-

stup projektu reálne ujal, potrebujeme posilniť digitálne kompetencie ľudí, ktorí s ním budú pracovať. V projekte preto počítame so vzdelávaním pracovníkov archívov aj verejnej správy. Vzdelávanie nebude izolovaný seminár, ale súčasť každodennej praxe.

Archívári sa budú učiť priamo pri testovaní na reálnych dokumentoch. Cieľom je, aby po skončení projektu nezostal v archíve „čierny box“, ale tím ľudí, ktorí mu rozumujú a vedia ho ďalej rozvíjať. Kvalitné ľudské zdroje sú zárukou udržateľnosti projektu.

Pri vzdelávaní sa črtá možnosť užšej spolupráce s Univerzitou Pavla Jozefa Šafárika, ktorá má rozsiahle skúsenosti so vzdelávaním v oblasti rozvoja digitálnych kompetencií. Projektu hrá do karát nový vysokoškolský zákon, pretože zavádza mikroosvedčenia a krátke študijné programy.

Ak sa projekt podarí, ako môže ovplyvniť slovenské archívy v horizonte najbližších rokov?

Ak pracovný tok overíme v podmienkach Štátneho archívu v Košiciach, získame model, ktorý je možné prispôsobiť iným archívom na Slovensku a potenciálne aj v zahraničí. Stavíme na medzinárodných štandardoch, no zároveň reflektujeme špecifiká Slovenska – od historického vývoja, cez štruktúru archívnej siete, až po spôsob, akým funguje verejná správa.

Dôležité je, že nejde len o technológiu, ale o zmenu organizácie práce. Časť rutinných činností, ktoré dnes zaberajú väčšinu dňa – prekladanie papierov, prepisovanie údajov či manuálne vyhľadávanie – môže postupne nahradiť automatizácia.

Odborný personál v archíve sa tak bude môcť viac sústrediť na činnosti s vyššou pridanou hodnotou: okrem samotnej validácie dát napríklad na metodické usmerňovanie, hodnotenie archívnej hodnoty doku-

mentov, výskum, popularizáciu fondov, prácu s verejnosťou či školami a pod. To priamo prispieva k personálnemu rozvoju archívárov. Namiesto „nosičov škatúl“ a operátorov skenera sa z nich stávajú odborní garanti digitálneho archívnictva, bádatelia a inovátori.

Z dlhodobého hľadiska ide o krok k modernému archívnictvu, kde sa archívne dedičstvo neukladá len „do škatule“, ale cieľavedome sa uchováva a sprístupňuje v digitálnej podobe. Časť minulosti pravdepodobne už nezachrátime, pretože materiál fyzicky degraduje a dlh sa nedá vymazať zo dňa na deň.

Môžeme však vytvoriť systém, ktorý prestane problém odkladať a začne ho skutočne efektívne riešiť. Projekty ako ArchAlve ukazujú, že technológie, výskum a prax sa dajú spojiť tak, aby z toho profitovalo kultúrne dedičstvo, samotní archívári aj celá spoločnosť.

Prečo by digitalizácia archívov mala zaujímať bežného človeka?

Na prvý pohľad sa môže zdať, že archívy sú skladmi starých dokumentov pre historikov. V skutočnosti sa ich fungovanie môže dotknúť každého, kto si vybavuje dôchodok, potrebuje kópiu strateného vysvedčenia, rieši dedičstvo alebo sa zaujíma o minulosť svojej rodiny.

Archívny dlh sa dnes prejavuje tak, že mnohé dokumenty sú buď fakticky nedostupné, alebo vyhľadávanie v nich stojí nemalé úsilie. Do archívu musíte prísť spravidla osobne, čakať, kým archívár dokumenty nájde, a ani nemáte vopred istotu, že nájdete, čo potrebujete.

Riešením je digitalizácia, ale sama o sebe už dnes nestačí. Ak len oskenujeme strany, máme síce „obrázok“, ale nie komplexnú informáciu. Kľúčové sú metadáta, teda štruktúrované údaje o dokumente. Ak sa podarí náš pracovný tok realizovať v praxi, bude možné cielenie vyhľadávať v dokumentoch podľa vybraných kritérií a veľkú časť dokumentov bude možné sprístupniť na diaľku.

Metadáta sú rovnako kľúčové pre výskum. Historik nemusí stráviť týždne prezeraním jednotlivých dokumentov v škatuliach. Vďaka metadátam dokáže rýchlo zistiť, ktoré dokumenty sú relevantné, a namiesto nekonečného hľadania dokumentov sa môže venovať ich analýze a interpretácii. Učiteľ dejepisu si cez elektronický inventár jednoducho vyhľadá dôležité pramene k učivu a môže skvalitniť svoju výučbu.

Digitalizácia archívov spolu s kvalitnými metadátami teda nie je luxus pre úzky okruh odborníkov. Je to investícia do funkčnej verejnej správy, právnej istoty, transparentnosti a živej historickej pamäti. Profituje z nej celá spoločnosť.

ELENA DANKOVÁ