Komplexné spracovanie písomného kultúrneho a vedeckého dedičstva SR

Martin Katuščák
Žilinská univerzita, Fakulta humanitných vied, Katedra mediamatiky a kultúrneho dedičstva
martin.katuscak@mediamatika.sk
Miroslav Čapkovič
Slovenská národná knižnica v Martine
miroslav.capkovic@snk.sk

INFORUM 2012: 18. konference o profesionálních informačních zdrojích
Praha, 22. - 24. 5. 2012

Abstrakt:

Práca sa zaoberá výskumno-vývojovými perspektívami národného projektu Digitálna knižnica a digitálny archív zameraného na masovú digitalizáciu a analógové a digitálne uchovanie písomných knižničných a archívnych materiálov, v ktorom budú implementované výsledky výskumu najlepších dostupných technológií a metód, pričom bude uplatňovaný integrovaný prístup ku spracovaniu fyzických a digitálnych foriem. Okrem uloženia kníh na bezpečné a čisté miesto, uľahčenia prístupu k plným textom a prezerania digitálnych obrazov existuje veľký potenciál skrytý v možnostiach ďalšieho spracovania digitálneho textu a jazyka, objavovania znalostí v textoch, prepájania obsahu a pod., ktoré sú zamerané na uvádzanie nových služieb a zlepšenie pracovného toku spracovania a jeho kroky, akými sú klasifikácia dokumentov, OCR, tvorba metaúdajov atď., ktoré vyžadujú ďalší výskum a metodologickú prácu založenú na spolupráci medzi príslušnými zainteresovanými stranami, ako sú pamäťové inštitúcie, výskumné pracoviská, vysoké školy, výrobcovia technológií a, samozrejme, používatelia. Príspevok poskytuje opis súčasných a plánovaných procesov, ako aj informácie o súvisiacich aktivitách vo výskumno-vývojovom projekte Pamäť Slovenska.

Úvod

Pre inštitúcie z oblasti kultúrneho a vedeckého dedičstva, ako sú archívy, knižnice, múzeá, galérie a i., vzniká nová situácia, vyžadujúca nové rozhodnutia a organizačné opatrenia na dlhodobé uchovávanie a sprístupňovanie nových digitálnych formátov a prinášajúca nové výzvy v súvisiacom výskume a vývoji. Knižnično-informačná veda má za cieľ prinášať interdisciplinárny pohľad na problematiku komunikácie, uchovávania a sprístupňovania informácií a poznatkov, pričom najrozšírenejším spôsobom kódovania signálu na prenos informácií je písmo, ako systém znakov charakteristický pre konkrétny jazyk a jeho súvisiacu kultúru. Ľudstvo počas dejín vytvorilo množstvo znakov v stovkách rozličných znakových sústavách a rôznych materiáloch, z ktorých sa zachovala len malá časť, pretože množstvo

efemérneho textu zaniklo ihneď po splnení komunikačného účelu. Množstvo zachovaných materiálov tvorí kolektívna pamäť ľudského spoločenstva, ktorú je nevyhnutné uchovávať pre ďalšie generácie v záujme zachovania kontinuity myslenia a prežitia ľudstva.

Existuje predpoklad, že dobre premysleným a cieľavedomým kvalitným spracovaním existujúcich a priebežne vznikajúcich poznatkov (významov) zaznamenaných akýmikoľvek znakmi – konkrétne ako text, je možné objavovať nové znalosti a vytvárať tak novú merateľnú hodnotu. Je nevyhnutné uchovávať a pokračovať v kontinuite komunikácie poznatkov kultúrneho a vedeckého dedičstva konkrétneho etnika a historicko-geograficky vymedzeného teritória – Slovenska za synergie s ostatnými slovanskými aj neslovanskými národmi a ich inštitúciami, ktoré čelia rovnakým výzvam vyplývajúcim z úloh zhromažďovania, spracovania, sprístupnenia a dlhodobé uchovania dokumentov vo fyzickej forme ako aj v digitálnej forme.

Situácia v Slovenskej republike

V prípade Slovenskej republiky prináleží plnenie úloh zhromažďovania, ochrany, uchovania a sprístupnenia písomného kultúrneho a vedeckého dedičstva (KaVD) knižničnému systému riadenému Slovenskou národnou knižnicou ako štátnou inštitúciou Ministerstva kultúry a archívnemu systému riadenému Ministerstvom vnútra – Sekcii archívnej správy (Slovenský národný archív). Ide o ústredné subjekty, ktoré podľa zákona zabezpečujú zhromažďovanie, spracovanie, ochranu, dlhodobé uchovanie a sprístupnenie písomných materiálov kultúrneho a vedeckého dedičstva a metodicky koordinujú ďalšie relevantné organizácie v rozsahu ich pôsobnosti. Tieto dve inštitúcie sú realizátormi národného projektu Digitálna knižnica a digitálne múzeum v rámci 2. prioritnej osi Rozvoj pamäťových inštitúcií a ich infraštruktúry Operačného programu Informatizácia spoločnosti (OPIS, 2007-2013), ktorý má s rozpočtom 49,6 mil € do roku 2015 spracovať 2,8 milióna objektov kultúrneho dedičstva, čo predstavuje približne 270 miliónov strán.

V rámci identifikácie výskumno-vývojových požiadaviek Slovenskej národnej knižnice v súvislosti s realizáciou národného projektu Digitálna knižnica a digitálny archív boli určené tri základné oblasti záujmu:

- Procesy súvisiace s ochranou a uchovávaním analógových nosičov a ich obsahu
- Integrácia procesov analógového a digitálneho spracovania
- Procesy súvisiace s digitalizáciou a ďalším spracovaním textu a ich optimalizácia

Slovenská národná knižnica a Slovenský národný archív sú povinné implementovať výsledky projektu štátnej výskumnej úlohy KNIHA SK – Záchrana, stabilizácia a konzervácia tradičných nosičov informácií v SR.

Národný projekt Digitálna knižnica a digitálny archív má celkový rozpočet 49,6 milióna € a plánuje vyprodukovať 270 miliónov strán, čo predstavuje náklady na jednu stranu približne 0,18 €. Okrem vyššie uvedených procesov tento projekt počíta okrem digitalizácie aj s vykonaním úkonov na ochranu analógových nosičov vybudovaním kapacít na masovú sterilizáciu a deacidifikáciu, čiže náklady zahŕňajú kompletné očistenie, chemické ošetrenie, reštaurovanie, logistiku atď.

Pri existujúcom stave poznania je možné uskutočniť komplexné spracovanie písomného kultúrneho a vedeckého dedičstva Slovenskej republiky, ktorého výsledkom bude vytvorenie merateľnej hodnoty v podobe nových znalostí. Ako príklad vytvorenej hodnoty možno uviesť výrok Swansona (1988)[1], podľa ktorého je možné „nové znalosti získať z knižnice rovnako ako z laboratória“ a je potrebné „stavať sa k informáciám rovnako ako k samotnému výskumu."

Písomné kultúrne a vedecké dedičstvo

Písomné kultúrne a vedecké dedičstvo je archívna a bibliografická, umelecko-historická a vedecká hnuteľná hmotná pamäť ľudského spoločenstva zaznamenaná v písomnej forme, ktorej existencia vyžaduje ochranu pred zánikom, a ktorej účel sa plní v interakcii s človekom – používateľom.

Jednotkou písomného kultúrneho a vedeckého dedičstva sú objekty, t.j. ucelené súbory údajov, informácií alebo poznatkov komunikované v zaužívaných formách ako sú monografie, noviny, časopisy, vedecké zborníky, kvalifikačné práce, výskumné správy, patenty a podobne. Písomné kultúrne a vedecké dedičstvo môže byť vytvorené prepisom iných foriem obsahu – napríklad prepis hovoreného slova, titulky k audiovizuálnemu dielu, textová anotácia pre iné ako písomné formy kultúrneho a vedeckého dedičstva a podobne.

Pri komplexnom spracovaní písomného kultúrneho a vedeckého dedičstva ide o široký okruh činností od identifikácie objektu, jeho získania, evidencie, katalogizácie, pripísania metaúdajov, konverzie do rôznych foriem a dlhodobého uchovania v týchto formách, ochrany a logistického zaradenia nosiča, až po doručenie koncovému používateľovi podľa požiadavky. Nové technológie prinášajú do knižničnej a archívnej praxe úplne nové metódy a postupy. Napriek výhodám, ktoré prinášajú nové digitálne formáty, z pohľadu knihovníkov a archivárov sa nemení situácia s analógovými dokumentmi, ktoré sú naďalej ohrozené vzhľadom na degradáciu materiálov nosičov. Komplexné spracovanie sa teda na jednej strane zameriava na **analógové formáty** (nosiče) písomného KaVD, čiže ide o procesy vzťahujúce sa najmä na ochranu a konzervovanie ako sú čistenie, reštaurovanie,

[1] Swanson, D. R. (1988). Historical note: Information retrieval and the future of an illusion. In Journal of the American Society for Information Science, Vol. 39, pp 92-98.

rekonštrukcia, sterilizácia, deacidifikácia, lyofilizácia knižničných a archívnych nosičov. Patria sem aj procesy preformátovania, napríklad ochranné fotokopírovanie alebo produkcia kópií na mikroformách.

Na druhej strane sa komplexné spracovanie zameriava na **digitálne formáty**, ktoré poskytujú jednoduchý a rýchly prístup k náhrade obsahu na analógových nosičoch, pričom ich životnosť je krátkodobá a zachovanie obsahu je podmienené opakujúcimi sa investíciami do aktualizácií na prekonanie technologického zastarávania. Digitálne formáty vyžadujú zložité fyzické nosiče, ktoré sa stávajú predmetom ochrany rovnako ako analógové nosiče.

Uvedené úlohy knižníc a archívov v oblasti ochrany fyzických nosičov sú v praxi pomerne jednoduché.

Z pohľadu uplatnenia vedeckého prístupu a nevyhnutnosti výskumu a vývoja je nosnou témou počítačové spracovanie KaVD, od ktorého sa očakáva nielen inteligentná organizácia poznania a podpora všetkých procesov komplexného spracovania, ale aj vytvorenie príležitostí na objavovanie nových znalostí. Podľa C. Sporleder (2010) je na odhalenie poznatkov nachádzajúcich sa vo fondoch a zbierkach kultúrnych inštitúcií potrebné písomné dedičstvo „očistiť, prepojiť a obohatiť“[2] Podľa Stephena M. Griffina (2010) je možné vytvoriť „digitálnu reprezentáciu pre množstvo zdrojov kultúrneho dedičstva“ za použitia počítačového spracovania, ktorý to vníma ako „prostriedok, ako obnoviť to, čo bolo stratené. Počítačové spracovanie pomocou geopriestorových a časových údajov je ústredným bodom vizualizácie a pochopenia mechanizmov zmeny v priebehu dlhých časových období“. Sú to práve nové počítačové technológie, ktoré umožňujú spracovanie obrovského množstva surových údajov a ich kombinovania na tvorbu nových informácií a poznatkov, za realizácie „kreatívnych prístupov, imaginatívneho myslenia a medzinárodnej interdisciplinárnej spolupráce“[3].

Pri komlexnom spracovaní je riadený celý cyklus spracovania materiálov - od bezkontaktnej rádiofrekvenčnej identifikácie objektu, jeho sledovania a kontroly v procesoch pasportizácie, ochrany, deacidifikácie, digitalizácie, mikrofilmovania, digitálneho spracovania, publikovania, vytvorenia digitálneho záznamu v digitálnej knižnici alebo digitálnom archíve, s prepojením na príslušné metaúdaje.

Vymedzenie ďalších súvisiacich pojmov (digitalizácia, digitálne uchovávanie, dolovanie znalostí z textov) je súčasťou prehľadu problematiky.

---

[2] Sporleder, C. (2010), Natural Language Processing for Cultural Heritage Domains. Language and Linguistics Compass, 4: 750–768. doi: 10.1111/j.1749-818X.2010.00230.x

[3] Griffin, Stephen M. 2010. Recovering the past through computation: new techniques for cultural heritage. In Proceedings of the international conference on Multimedia information retrieval(MIR '10). ACM, New York, NY, USA, 13-14. DOI=10.1145/1743384.1743394 http://doi.acm.org/10.1145/1743384.1743394

Problematika komplexného spracovania zahŕňa nasledujúce oblasti:

- **Práca s analógovými materiálmi písomného KaVD** (knižničná a archívna prax, integrácia nových digitálnych technológií).
- **Digitalizácia písomného KaVD.**
- **Získavanie obrazu a textu – optické rozpoznávanie znakov** (s využitím výstupu z ďalšieho spracovania textu).
- **Práca s dokumentami a textom** - analýza textu - jazyková analýza, štrukturálna analýza atď., budovanie jazykových korpusov, dolovanie a objavovanie znalostí v texte, bibliometrické a štatistické analýzy.
- **Metaúdaje** – na opis, sprístupnenie, uchovanie, a riadenie komplexného spracovania, právne metaúdaje.
- **Použitie a použiteľnosť** – spôsob sprostredkovania obsahu písomného KaVD pomocou tradičných a najmodernejších médií.

V rámci **integrácie komplexného spracovania fyzických a digitálnych foriem** je potrebné klásť dôraz na vzájomnú synergiu úkonov fyzickej ochrany (uchovávania) materiálov písomného KaVD procesmi digitálneho spracovania používaním vhodných metaúdajov, plánovaním a zaznamenávaním a realizáciou stratégií ochrany fyzických nosičov, využívanie najmodernejších metód na organizáciu a lokalizáciu fyzických predmetov písomného KaVD (logistické postupy, rádiofrekvenčná identifikácia, geografické informácie atď.).

Digitalizácia je rozsiahly proces pozostávajúci z viacerých vrstiev a krokov. Samotný prevod na sériu núl a jednotiek kódujúcich body obrazu alebo znaky textu je len jeho malou časťou. V prípade písomných materiálov začína digitalizácia identifikáciou a výberom materiálu, jeho posúdením a prípravou pre vhodné technologické zariadenie, pokračuje vytvorením digitálnych obrazov a následuje úprava obrazov.
V podmienkach Slovenskej republiky bude samotná digitalizácia v rámci národného projektu Digitálna knižnica a digitálny archív realizovaná na profesionálnej úrovni ako masová digitalizácia alebo „digitalizácia vo veľkom meradle“ (angl. large-scale digitisation), najmä pomocou digitalizačných robotov s automatizovaným obracaním strán v prípade zachovalých viazaných materiálov, vysokorýchlostných prietokových skenerov na voľné listy a krehkejšie dokumenty budú na základe precíznej selekcie nasmerované na manuálnu digitalizáciu.
Digitalizácia písomného kultúrneho a vedeckého dedičstva predstavuje v podmienkach SR proces konverzie analógových materiálov ako sú knihy, časopisy, historické knižné dokumenty, rukopisy, vyhľadávacie pomôcky, inventáre, archívne spisy, čiže materiálov textovo-vizuálneho charakteru povahy do digitálnej formy vhodnej na dlhodobé digitálne uchovanie a sprístupnenie obsahu – údajov, informácií, poznatkov.

V rámci spoločnej stratégie digitalizácie kultúrneho, vedeckého a intelektuálneho dedičstva má knižničný a archívny systém zodpovednosť najmä za písomné dedičstvo, ktoré bude chránené proti degradácii materiálov nosičov a následnej strate obsahu a digitalizované uplatnením priemyselných postupov.
V súčasnosti sa problematike zlepšovania prístupu k textu venuje projekt financovaný Európskou komisiou s názvom IMPACT[4], ktorý v roku 2011 predstavil Európskemu spoločenstvu model kompetenčného centra a sadu nástrojov potrebných na plánovanie a realizáciu masového spracovania a sprístupnenia textu, vrátane historických dokumentov www.impact-project.eu a www.digitisation.eu.

Skúsenosti z veľkých projektov digitalizácie a metodické odporúčania pre digitalizáciu ukazujú, že digitalizácia nie je cieľ a koniec spracovania písomného kultúrneho a vedeckého dedičstva, ale len predspracovanie dokumentov na ďalšie účely ako sú rozpoznávanie znakov, štrukturálna analýza, dolovanie znalostí z textov. Ide o nevyhnutný predpoklad na všetky následné procesy so získaným digitálnym textom.
Výstupom digitalizácie knižničných a archívnych materiálov budú pôvodné a upravené digitálne obrazy strán a automaticky konvertované obrazy na text, s úspešnosťou rozpoznania správnych znakov približne 95%, ktorá je však nepostačujúca na účely ďalšieho spracovania textu a znižuje aj kvalitu ďalších poskytovaných služieb (napr. nepresnosť vyhľadávania). Úlohou výskumu je dosiahnuť 100 % úspešnosť rozpoznania znakov a celých slov z pôvodneho obsahu. Vyššia úspešnosť pri rozpoznavaní znakov závisí okrem kvalitného zosnímania predlohy aj od správnej identifikácie štruktúry textu, od odlíšenia textových informácií od obrazových informácií, kvality doménovo orientovaných slovníkov odvodených z korpusu slovenskeho jazyka. Špecializované slovníky (lexikony) sa priebežne aktualizujú o novorozpoznané slová a následne sa opäť využívajú v procese OCR a v nástrojoch sprístupňovania digitálneho obsahu. Cieľom výskumu a vývoja v tejto oblasti je priebežné dopĺňanie a organizácia pôvodných doménovo orientovaných a historicky segmetovaných slovníkoch používaných v procese OCR o jazykové jednotky, ktoré je možné získať len prostredníctvom masovej digitalizácie celého fondu slovacikálneho písomného kultúrneho a vedeckého dedičstva a prostredníctvom uvedeného procesu skvalitniť rozpoznávanie znakov a textu na úroveň 100 %. Predpokladá sa, že zdigitalizované obrazy budú musieť prejsť procesom OCR viackrát, pričom percento správne rozpoznaných znakov bude po každom procese vyššie, až kým nebude text rozpoznaný bezchybne. Počet opakovaní krokov bude závisieť od procesu obohacovania slovníkov. Aktualizácia slovníkov je množina lingvistických analýz, z ktorých bude musieť byť časť vykonaná manuálne.

[4] IMPACT Project: Improving Access to Text [webové sídlo] Dostupné na internete: <http://www.impact-project.eu/

Kolaboratívna korektúra textu je realizovaná prostredníctvom webovej služby, ktorá sprístupňuje korektorom originálny obraz strany dokumentu vo forme upravovateľného textu vygenerovaného automatickým procesom OCR. Zmenami, ktoré korektor aplikuje do textu, vznikajú nové vrstvy textu (verzie). Korektor môže v originálnom texte opraviť preklepy alebo technické chyby, avšak pridanou hodnotou takejto korektúry je analytická činnosť, pri ktorej korektor zistí v texte nové slovo alebo nový tvar slova a pridáho ho do vytváraného slovníka, prípadne pridá nové pravidlo alebo vzťah do korpusu. Korektori môžu zároveň podporovať proces automatického rozpoznávania vlastných mien. Korektúra je ako manuálny proces spojená s predspracovaním textu, keď je súčasne upravený a overený slovník použitý pred korektúrou na spracovávaný text. Samotnú lingvistickú prácu so slovníkmi a overovanie a schvaľovanie prírastkov do slovníkov a korpusov budú realizovať jazykovední experti a nimi vyškolení operátori. Pri uvedenom spracovaní sú aplikované metódy analýzy textu zamerané na novovytváraný korpus generovaný automatizovane a prostredníctvom manuálnych korektúr a na novovytváraný slovný fond. Výstupom spracovania budú časovo diferencované morfologické a lexikálne slovníky a autority a upravený časovo diferencovaný jazykový korpus. Priebežné dopĺňanie korpusu a slovníkov je jedným z najpodstatnejších aspektov zvýšenia kvality textu pôvodných materiálov a presnosti jeho rozpoznania, čo je nevyhnutný predpoklad na ďalšie procesy práce s textom.

Problematika metaúdajov súvisí aj s použitím nástrojov na organizáciu poznania akými sú názvoslovia, klasifikačné systémy, taxonómie a ontológie na vyjadrenie explicitných znalostí. V kontexte vznikajúceho Sémantického webu a vzhľadom na možnosti nástrojov Webu 2.0 sa pri tvorbe metaúdajov čoraz viac kladie dôraz na používateľa, ktorý je tiež zapájaný do procesu ich tvorby. S problematikou metaúdajov a spracovania textu súvisí udržiavanie a doplňovanie súborov autorít, čiže rozpoznávanie vlastných mien (pomenovaných entít, angl. Named Entity Recognition), tvorba a údržba a využívanie tezaurov a klasifikácií.

Pri spracovaní jazyka a textu sa v kontexte komplexneho spracovania kultúrneho a vedeckého dedičstva SR zaoberáme jeho hmotne zachytenou podobou – písmom, textom.

Fázy analýzy textu sú podľa Furdíka (2008)[5]nasledovné:

1. Konverzia textu na jednotný formát a extrakcia metaúdajov
2. Segmentácia a tokenizácia
3. Lematizácia a morfologická analýza
4. Slovotvorná analýza.
5. Syntaktická analýza.
6. Sémantická a pragmatická analýza.

[5] Furdík, Karol (2008), Algoritmy predspracovania textu pre úlohy klasifikácie a zhlukovania v systéme elektronickej výučby, Technická univerzita v Košiciach, 2008. Dostupné na internete: <http://web.tuke.sk/fei-cit/furdik/publik/Kolokvium07_furdik_2008_PoZnaT.pdf>

Proces spracovania jazyka je možné vďaka technologickému pokroku realizovať automatizovane, pričom ako najvhodnejšia jednotka jazyka je na účely spracovania slovo a hlavnými okruhmi jazykovej analýzy textu sú slovná zásoba, gramatika a význam.

Na Slovensku sa práci s jazykovými korpusmi venuje Jazykovedný ústav Ľudovíta Štúra[6], ktorý od roku 2004 poskytuje prístup ku Slovenskému národnému korpusu na výskumné a vzdelávacie účely. Národný korpus je „elektronická databáza obsahujúca slovenské texty z rôznych štýlov, žánrov, vecných oblastí, regiónov a pod., vybavená výkonným vyhľadávacím systémom a prídavnými jazykovými informáciami“ a jej aktuálna verzia obsahuje 719 miliónov tokenov. Slovenský národný korpus môže byť použitý ako zdroj jazykových informácií a slovníkov pre optické rozpoznávanie znakov a slov pri spracovaní KaVD a naopak, pri spracovaní KaVD bude korpus priebežne dopĺňaný o validované vstupy.

Predpokladá sa, že na dosiahnutie lepších výsledkov pri rozpoznávaní textu je vhodnejšie namiesto univerzálneho korpusu využívať doménovo-orientované korpusy špecializované na oblasti ako sú technika, prírodné vedy, medicína, právo, obchod atď.

Kombinovanie doménových korpusov pri objavovaní znalostí v texte môže zas priniesť nové prepojenia a súvislosti (napr. matematika a umenie).

Paralič (2010)[7] rozdeľuje a opisuje nasledujúce jednotlivé kroky dolovania znalostí z textov:

- pochopenie aplikačnej domény,
- získanie relevantnej množiny dokumentov,
- predspracovanie,
- dolovanie v textoch,
- vizualizácia a interpretácia výsledkov.

Používanie a použiteľnosť

Konečným cieľom a zmyslom spracovania písomného kultúrneho a vedeckého dedičstva je jeho sprístupnenie rozličným skupinám používateľov – od detí, cez učiteľov až po vedcov prostredníctvom špecializovaných a personalizovaných služieb – od fyzických výpožičiek až po aplikačné sieťové služby. Hypertext priniesol nové možnosti horizontálneho prepojenia textov a nesúvislého asociatívneho čítania v sieťovom kontexte, na ktorý je dnešný používateľ internetu zvyknutý. Popri zaužívaných spôsoboch percepcie textu ako je čítanie tlačeného textu z papiera alebo zobrazeného textu z obrazovky prinášajú nové technológie neustále nové spôsoby a možnosti interakcie. Dá sa predpokladať, že používateľská komunita bude vyžadovať dostupnosť a použiteľnosť možností objavujúcich sa pri

[6] Slovenský národný korpus / Jazykovedný ústav Ľudvíta Štúra Slovenskej akadémie vied [online] [cit. 12-2-2012] Dostupné na internete: <http://korpus.juls.savba.sk/>

[7] Paralič, J. et al. (2010) Dolovanie znalostí z textov. Košice : Technická univerzita, 2010. ISBN 987-80-89284-62-7 [cit. 14-01-2012] Dostupné na internete: <people.tuke.sk/jan.paralic/knihy/DolovanieZnalostizTextov.pdf>

implementácii nástrojov Webu 2.0, ako sú obohatenie obsahu o podporný kontext, podpora viacjazyčnosti, umožnenie vyhľadávania pôvodného obsahu v inom jazyku, preklade do iných jazykov, poskytovanie sieťových odkazov na existujúce poznatky. Tieto informácie sú generované priamo z pôvodného textu, ktorý nie je zmenený, ale doplnený v spojení s externými zdrojmi údajov, informácií a poznatkov.
Najväčší potenciál komplexného spracovania KaVD spočíva v príležitosti na objavovanie a kombinovanie nových znalostí s použitím najmodernejších nástrojov s umelou inteligenciou pri počítačovom spracovaní.

Optimalizáciou procesov komplexného spracovania KaVD v projekte Digitálna knižnica a digitálny archív je možné preukázať, že náklady na spracovanie môžu byť podstatne nižšie ako ceny v podobných projektoch vo svete. Aj napriek prepokladaným pozitívnym ekonomickým ukazovateľom stále existuje priestor pre optimalizáciu komplexného spracovania KaVD. Simuláciami technologických procesov bolo preukázané (Čapkovič, 2010)[8], že pre dané veľké množstvo digitalizovaného materiálov má „ľubovoľné malé zlepšenie technológie veľký vplyv na konečný efekt". Ide predovšetkým o technológie na optimalizáciu logistiky toku papierových foriem a digitálneho obsahu v rôznych štádiách spracovania.
Cieľom súvisiaceho výskumu je stanovenie pravidiel a zmien v metodologických a technologických postupoch na zefektívnenie životného cyklu papierovej a digitalnej formy a dosiahnutie čo najvyššej priepustnosti. Prostredníctvom spolupráce s technologickými výrobcami softvéru pre digitalizačné zariadenia, na úpravu obrazov, OCR je potrebné integrovať softvérové komponenty do komplexného pracovného toku so špecifikovanými pravidlami triedenia materiálov.

Záver

V podmienkach Slovenskej republiky je uskutočniteľný integrovaný a holistický prístup ku spracovaniu analógového a digitálneho písomného a kultúrneho a vedeckého dedičstva. Väčšinu procesov, najmä pri práci s digitálnymi obrazmi a digitálnym textom je možné z veľkej časti automatizovať, stále však zostávajú procesy, ktoré je nutné vykonávať manuálne. Pripravovaný národný projekt Digitálna knižnica a digitálny archív je jedinečnou príležitosťou aplikovať vedecký prístup pri definovaní pracovných postupov a riešení praktických problémov, prostredníctvom spolupráce vedecko-výskumných a akademických pracovísk na Slovensku s praxou realizovanou v relevantných organizáciách, ktoré spravujú písomné dedičstvo a všetky podoby jeho nosičov tak, aby mohlo byť s pridanou hodnotou

[8] Čapkovič, Miroslav (2010). Digitalizácia v Slovenskej národnej knižnici. Prezentácia na konferencii Digitálna knižnica 2010. Demänovská dolina - Jasná

odovzdávané budúcim generáciám. Napriek koordinovanému postupu a výmene znalostí v rámci medzinárodného spoločenstva je plnenie tejto povinnosti v zodpovednosti každej krajiny. Záleží na intenzite spolupráce zainteresovaných strán, do akej miery bude možné vyťažiť potenciál z pôvodného kultúrneho a vedeckého obsahu.

Zoznam bibliografických odkazov

ČAPKOVIČ, Miroslav. 2010. *Digitalizácia v Slovenskej národnej knižnici*. Prezentácia na konferencii Digitálna knižnica 2010. Demänovská dolina - Jasná

ČAPKOVIČ, Miroslav. 2011., *Aktivity_WP* [dokument Microsoft Word], Interná elektronická komunikácia, Slovenská národná knižnica. 28-12-2011, 10 s.

FAN, W., et al. 2005. *Tapping Into the Power of Text Mining* [online] [cit 14-02-2012]. Dostupné na internete: <http://filebox.vt.edu/users/wfan/paper/text_mining_final_preprint.pdf>

FURDÍK, Karol. 2008. *Algoritmy predspracovania textu pre úlohy klasifikácie a zhlukovania v systéme elektronickej výučby* [online]. Košice : Technická univerzita v Košiciach, 2008. Dostupné na internete: <http://web.tuke.sk/fei-cit/furdik/publik/Kolokvium07_furdik_2008_PoZnaT.pdf>

IMPACT Project: *Improving Access to Text* [webové sídlo] Dostupné na internete: <http://www.impact-project.eu/

PARALIČ, J. et al. (2010) *Dolovanie znalostí z textov* [online]. Košice : Technická univerzita, 2010. [cit. 14-01-2012] Dostupné na internete: <people.tuke.sk/jan.paralic/knihy/DolovanieZnalostizTextov.pdf> ISBN 987-80-89284-62-7

Slovenský národný korpus [online] / Jazykovedný ústav Ľudvíta Štúra Slovenskej akadémie vied [cit. 12-2-2012] Dostupné na internete: <http://korpus.juls.savba.sk/>

GRIFFIN, Stephen M. 2010. *Recovering the past through computation: new techniques for cultural heritage* [online]. In Proceedings of the international conference on Multimedia information retrieval (MIR '10). ACM, New York, NY, USA, 13-14. DOI=10.1145/1743384.1743394 http://doi.acm.org/10.1145/1743384.1743394

SWANSON, D. R. 1988. *Historical note: Information retrieval and the future of an illusion*. In Journal of the American Society for Information Science, Vol. 39, pp 92-98.

SPORLEDER, C. (2010), Natural Language Processing for Cultural Heritage Domains. Language and Linguistics Compass, 4: 750–768. doi: 10.1111/j.1749-818X.2010.00230.x