NAUTILUS – Pomáhame sprístupniť informácie už raz nájdené

Nemusíte byť študentom, aby ste potrebovali nájsť nejaké informácie. My v škole pracujeme prevažne s digitálnymi knižnicami. S radosťou sme preto prijali správu o tom, že vďaka digitalizácii máme dnes len v Univerzitnej knižnici k dispozícii viac ako milión strán z periodík. Lenže o to väčšie bolo naše sklamanie, keď sme zistili, že vyhľadávanie v plných textoch je síce fajn, ale tie desiatky strán, ktoré získame, nie sú žiadnou výhrou. Chcelo by to analytický rozpis článkov. Lenže je to vôbec možné? Veď na jednej takej strane môže byť hneď niekoľko článkov. Tieto a podobné otázky sme si na začiatku kládli aj my. Rozhodli sme sa však využiť naše znalosti a skúsenosti s technológiami a výsledkom je ....

Obr. 1

Dôvody a východiská projektu

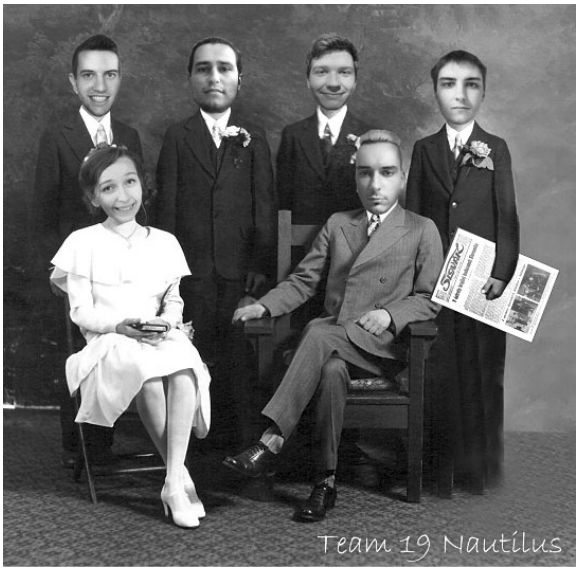
V dôsledku webu 2.0, ale aj rozvoja ďalších informačno-komunikačných technológií sa neustále zrýchľuje nárast dát, a to nielen v textovej podobe. Na druhej strane knižnice a iné inštitúcie tohto typu majú vo svojich fondoch obrovské množstvá dát, ktoré sú ešte aj teraz pre všetky technológie prakticky neprístupné (neviditeľné). Jedná sa predovšetkým o staršie periodiká, ktoré sú spracované len na úrovni súborného záznamu periodika. To, o čom sa píše v jednotlivých článkoch je však pre počítače úplne neviditeľné, aj keď v mnohých prípadoch ide o veľmi cenné informácie a poznatky. Vyhľadávať v nich môžeme len manuálne, teda môžeme čítať obsahy jednotlivých čísiel a konkrétne články a len tak môžeme nájsť tie informácie, ktoré práve potrebujeme alebo môžeme tak získať pre nás zatiaľ úplne nové poznatky. V dôsledku masívnej digitalizácie, ktorá sa u nás začala realizovať pred pár rokmi sa podarilo časť týchto zdrojov dostať do digitálnej podoby, takže sa stali aj pre počítače prístupné. Sprístupnenie a vyhľadávanie sú však dve úplne rozličné veci. Aj keď sú tieto dáta v digitálnej podobe, stále chýbajú nástroje, ktoré by umožňovali rýchlo, odkiaľkoľvek a z akéhokoľvek zariadenia nájsť to, čo práve potrebujeme. Oveľa lepšie je to s výberovými periodikami a aj to len z časti, pretože tu niektoré inštitúcie pristúpili k spracovaniu periodík aj na úrovni analytického rozpisu. To znamená, že bibliografický záznam, vytvárajú nielen pre „časopis" ako taký, ale tiež pre každé číslo periodika a výberovo aj pre určité články z daného čísla. Takže obsah týchto článkov je rovnako prístupný ako knihy, a to bez ohľadu na to, či sa jedná o klasickú formu dokumentu alebo je dokument vydaný v elektronickej podobe.

Preto sme sa rozhodli v našom projekte vyriešiť proces automatizácie spracovania analytického rozpisu tých periodík, ktoré boli zdigitalizované a bol na nich aplikovaný OCR program. Väčšina inštitúcií u nás k tomuto účelu používa sw nástroj známy ako Abby Recognition, ktorý takto spracované dáta ukladá v špeciálnom XML súbore danej verzie. Aplikácia, ktorá je výsledkom nášho projektu, pomáha knihovníkom pri spracovaní analytického rozpisu archivovaných periodík a tým pádom ich sprístupňuje širokému spektru používateľov. Pomáha taktiež vykonávať ďalšie úpravy textu, ktorý je výsledkom OCR tak, aby bolo možné s ním pracovať pri aplikovaní nástrojov umelej inteligencie a umožniť tak aj úplne presné vyhľadávanie na základe významu. Takto spracované periodiká umožnia bežnému používateľovi nielen rýchlejšie nájsť to, čo práve potrebuje, ale následne tiež identifikovať a odhaliť nové poznatky, ktoré tieto staré dokumenty ukrývajú.

Niečo o nás

Tím Nautilis, to sme my, mladí, motivovaní študenti Fakulty informatiky a informačných technológií STU v Bratislave a pod vedením Ing. Nadeždy Andrejčíkovej, PhD., sa snažíme využiť IKT tak, aby priamo pomáhali pri sprístupňovaní kultúrneho

dedičstva. Na tomto projekte sme tiež priamo spolupracovali s jednou z popredných našich inštitúcií v tejto oblasti, a to s Univerzitnou knižnicou v Bratislave, ktorá okrem iného spravuje rozsiahly fond tlačených periodických dokumentov a buduje Súborný katalóg periodík. Univerzitná knižnica v posledných rokoch zdigitalizovala veľké množstvo týchto starších ročníkov a čísiel periodík.

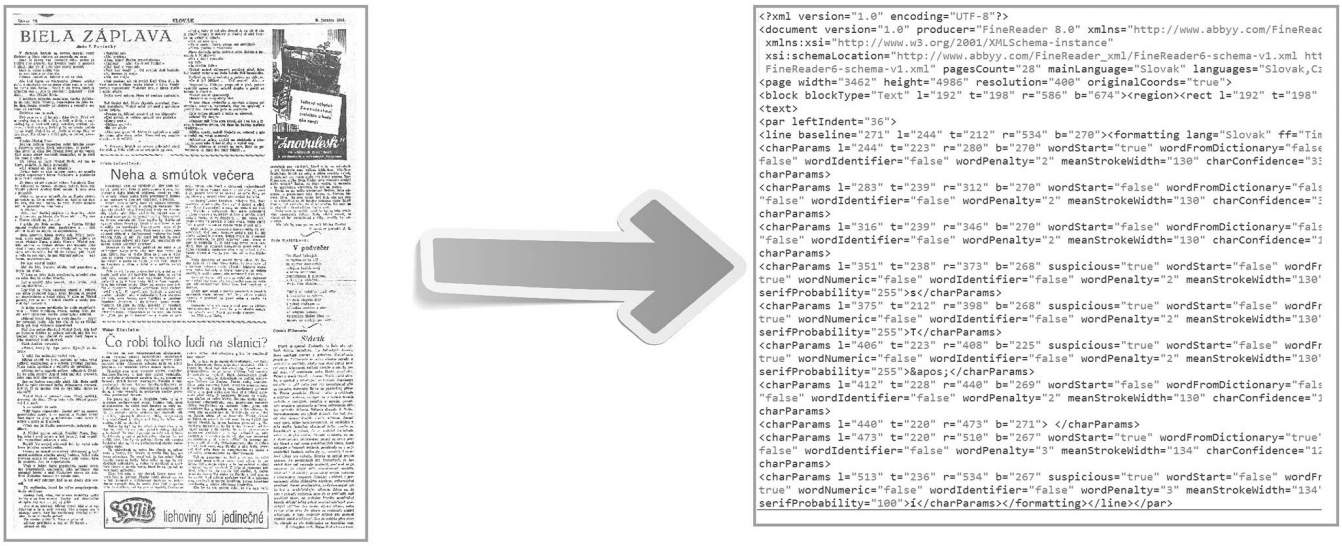


Obr. 2

Spracovanie zdrojov v súčastnosti

Ako sme už uviedli, súčasťou procesu digitalizácie, či skôr tzv. post procesingu, je rozpoznávanie znakov, k čomu sa využíva Adobe Recognition Software. Tento umožňuje zdigitalizované obrazy klasických dokumentov transformovať do špeciálnych XML dokumentov. Takéto súbory, ale aj bežné txt dokumenty, či dokumenty vytvorené v bežných textových editoroch sú tie, na ktoré sa zameriavame a ktoré tvoria základný vstup do nášho projektu.

Tieto špeciálne XML súbory síce obsahujú množstvo informácií, ale sú to zväčša informácie týkajúce sa formátovania a úpravy textu, čo pre účely vyhľadávania nemá dostatočný význam.



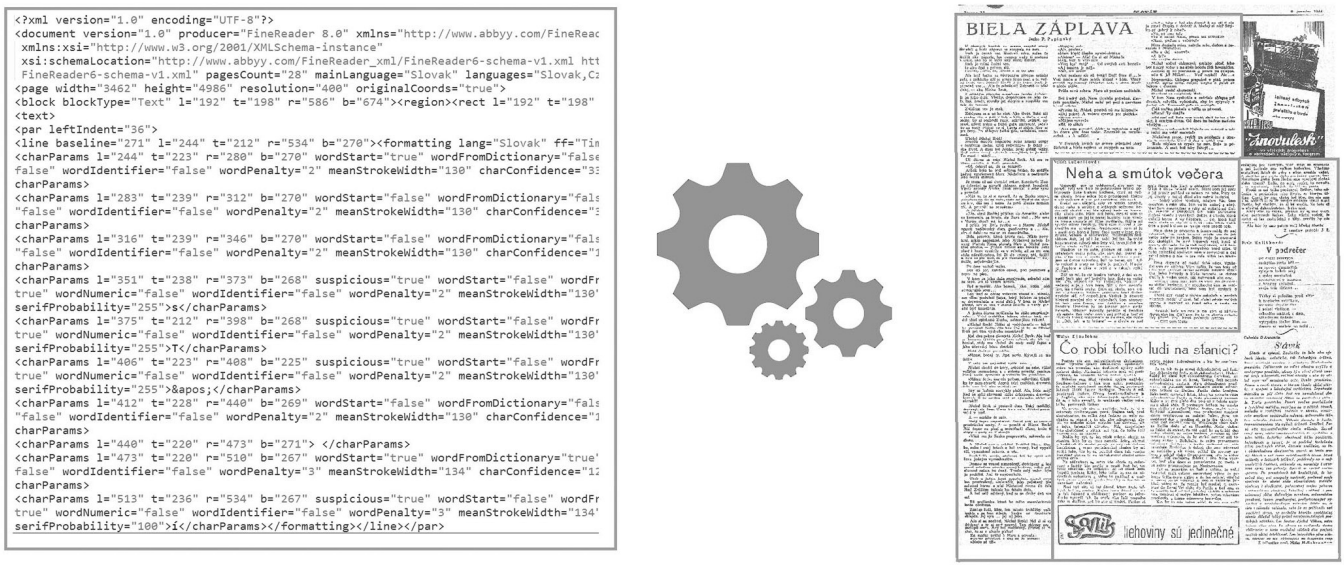
Obr. 3

Ako to robíme a náš cieľ

Našim hlavným cieľom v tomto projekte je, ako sme už uviedli, identifikovať tie informácie, ktoré nám umožnia správne rozpoznať jednotlivé bloky textov, ale aj to, ktoré a v akom poradí patria ešte k sebe a tiež to, ktorý z týchto blokov je ten, čo určuje názov, resp. názvovú informáciu. Ako to robíme?

Krok 1: predspracovanie článkov

V prvom kroku, skôr než sme vstupný súbor začali spracovávať, sme museli vykonať viaceré kontroly zamerané na kvalitu vstupného súboru. V tejto časti sme aplikovali viacero metód tak, aby sme mohli spracovateľovi, či administrátorovi poskytnúť čo najpresnejšiu informáciu o nezrovnalostiach, ktoré sme týmito analýzami identifikovali. V prípade, že vstupný súbor spĺňal kvalitatívne kritériá, sme pristúpili k predspracovaniu vstupného súboru. Cieľom tohto predspracovania uvedených vstupných súborov je dosiahnuť vo výslednej forme paragrafy jednotlivých článkov v rovnakých skupinách. K tomuto sme museli vykonať podrobnú analýzu týchto vstupných dát, ktoré sa najčastejšie nachádzajú v rôznych verziách. Identifikovali sme tu viaceré kvalifikátory a iné pravidlá, kombináciou ktorých sa nám podarilo zadefinovať základný algoritmus. Pomocou tohto nášho algoritmu môžeme automatizovať proces extrakcie článkov. Tento algoritmus je založený na rozpoznávaní nadpisov a textov a následnom priraďovaní jednotlivých textov k prislúchajúcim nadpisom. Vďaka našej metóde sme už dnes schopní odhaliť a extrahovať až 70 % článkov zo vstupných XML súborov jednotlivých čísel periodík.



Obr. 4

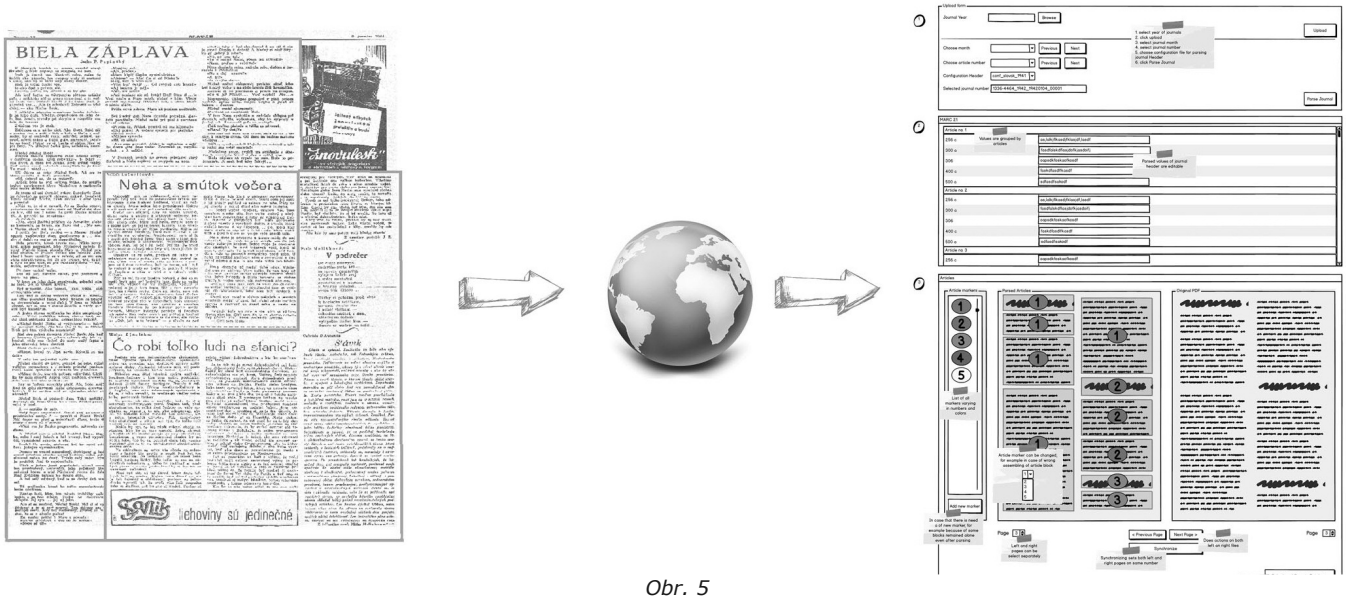
Následne sa snažíme tieto texty spracovať a identifikovať v nich kľúčové slová, neskôr aj význam, v akom boli tieto slová v danom texte použité. Keďže hovoríme o automatizácii procesu spracovania periodík na úroveň ich analytického rozpisu, tak z takto získaných údajov generujeme bibliografické záznamy pre dané číslo periodika, ako aj pre konkrétne články, ktoré sú v ňom obsiahnuté. Samozrejme všetky tieto nami generované bibliografické záznamy článkov obsahujú vždy väzbu na dané číslo periodika a bibliografický záznam čísla periodika, mjú v sebe väzbu na existujúci záznam súborného záznamu periodika, prípadne je možné pridať aj väzbu na záznam zviazaného ročníka a zväzku periodika, ale to by si vyžadovalo pri nastavení služby mať k dispozícii tieto bibliografické záznamy zviazaných ročníkov periodík. Bibliografické záznamy generujeme vo formáte MARC21/XML podľa platných katalogizačných pravidiel, aby s nimi mohli priamo pracovať aj knižnično-informačné systémy. K jednotlivým bibliografickým záznamom ukladáme do nášho repozitára aj plné texty a ich vizualizované obrazy tak, aby boli vzájomne prepojené a používateľ ich mohol získať spolu s bibliografickým záznamom.

Krok 2: webová aplikácia

Pridanou hodnotou celého nášho riešenia je priamo webová aplikácia, pomocou ktorej môžu používatelia jednoduchou a intuitívnou formou pracovať s výsledkami nášho riešenia, teda s jednotlivými rozpoznanými a identifikovanými textami a ich usporiadaním do článkov. Výhodou tejto webovej aplikácie je, že zároveň predstavuje rozhranie pre úpravu samotných plných textov. Keďže kvalita primárnych zdrojov, ako aj schopnosti OCR nástrojov rozpoznávať znaky, nemusí byť a z praxe vieme, že ani nie je vždy stopercentná, je súčasťou nášho projektu tiež návrh a realizácia aplikácie, ktorá umožňuje zamestnancom inštitúcií pre archiváciu takýchto zdrojov, prípadne iným používateľom, editovať aj výsledky OCR spracovania. Umožňuje im teda priamo opravovať daný text, ktorý pri opravách priebežne ukladáme. Rovnako, pomocou tejto aplikácie, umožňujeme používateľovi s danými právami aj upravovať výsledky nášho algoritmu a upresniť, či inak preorganizovať rozdelenie dokumentu na články, pretože vplyvom nekonzistencie tlače, a aj iných faktorov, nie je možné garantovať 100 %-nú úspešnosť rozpoznania článkov z týchto údajov.

Prínos tohto projektu

Výsledná aplikácia nášho projektu pomáha urýchliť a skvalitniť proces spracovania archivovaných periodických dokumentov na úroveň analytického rozpisu periodík a tým ich sprístupniť širokému spektru používateľov. Naše riešenie prináša metódu, ako automatizovať proces extrakcie konkrétnych článkov obsiahnutých v dokumente a zároveň pre každý z nich generovať plnohodnotný bibliografický záznam s väzbou na plný text článku a jeho vizualizovaný obraz. Vďaka výsledkom nášho projektu bude možné značne urýchliť a tiež následne skvalitniť spracovanie periodických dokumentov až na úroveň analytického roz-



Obr. 5

pisu periodík v tlačenej podobe. Toto následne vedie k novým možnostiam, ako aj ku skvalitneniu sprístupňovania týchto dokumentov a rovnako aj k novým možnostiam vyhľadávania, získavania a viacnásobného použitia týchto dokumentov. Kvalitný bibliografický popis týchto dokumentov môže viesť tiež k odhaľovaniu nových poznatkov, ktoré tieto dokumenty v sebe ukrývajú či už priamo, alebo v prepojení na iné dokumenty. Bežní používatelia tak získajú nový plnohodnotný zdroj širokého spektra informácií, v ktorých sa bude môcť štandardným spôsobom vyhľadávať a vybrané dokumenty aj priamo či sprostredkovane získavať. Dôležitý je fakt, že používateľ nebude nútený čítať, či inak vizuálne si spracovávať množstvo dokumentov a dlhé texty, ktoré teraz získava pri plnotextovom vyhľadávaní.

Realizácia tohto projektu je tiež krásnou ukážkou toho, že spoluprácou školy, inštitúcií štátneho či verejného sektora, spolu so súkromným sektorom je možné dosiahnuť pridanú hodnotu a výsledky, ktoré sú prospešné širokému spektru používateľov. Je to krásny príklad toho, ako nasmerovať aktivity pri spracovaní a sprístupňovaní kultúrneho dedičstva.

Možno by bolo vhodné sa nad tým zamyslieť, ako takúto spoluprácu naštartovať a následne aj aplikovať do bežnej praxe škôl a inštitúcií štátneho a verejného sektora, pretože naša krajina má unikátne kultúrne dedičstvo, ktoré predstavuje priam nevyčísliteľné bohatstvo. Správne formy jeho sprístupňovania, ako aj využívania môžu značne prispieť aj k rozvoju a ekonomickému rastu celej krajiny.

Bc. Martin Vaško vaskom1@stuba.sk
Ing. Andrejčíková Nadežda, PhD. nadezda.andrejcikova@stuba.sk ■

(Katedra knižničnej a informačnej vedy, Filozofická fakulta, Univerzita Komenského v Bratislave)