Súčasné trendy v digitalizácií a sprístupňovaní digitalizovaného obsahu v bratislavskej Univerzitnej knižnici

Autor sa v článku zaoberá digitalizáciou knižných dokumentov, manažmentom digitalizačných projektov, spracovaním digitalizovaných dát, sprístupnením digitalizovaného obsahu a jeho archiváciou. Predstavuje Odbor digitalizácie v bratislavskej Univerzitnej knižnici a informuje o jeho aktivitách, službách a procesoch.

Charakteristika pracoviska

Odbor digitalizácie Univerzitnej knižnice v Bratislave (ďalej len UKB) zabezpečuje operatívne knižnično-informačné digitalizačné služby pre odbornú verejnosť, širokú laickú verejnosť, pedagogických a vedeckých pracovníkov, pamäťové a fondové inštitúcie, študentov a mládež v celoštátnom a medzinárodnom rozsahu. Podieľa sa na digitalizácii a reštaurovaní unikátnych ohrozených a poškodených knižničných fondov. Zabezpečuje on-line doručovanie digitalizovaných dokumentov knižniciam a ich používateľom, manažuje realizáciu digitalizačných objednávok, digitalizáciu, digitalizačný postprocessing, archiváciu a sprístupňovanie digitalizovaných fondov literárneho kultúrneho dedičstva.

Obrázok č. 1 Odbor digitalizácie UKB.

Odbor digitalizácie UKB je prevádzkovaný v dvojzmennej prevádzke od 6.30 – 21.00 hod., v odbore digitalizácie pracuje 7 pracovníkov (6,75 úväzku), zastrešuje domáce a zahraničné digitalizačné projekty, koordinuje spoluprácu v oblasti digitalizácie a výmeny digitalizovaných knižničných dokumentov. Zabezpečuje riešenie úloh v súvislosti so sprístupnením a archiváciou digitalizovaného kultúrneho dedičstva. Univerzitná knižnica v Bratislave realizuje digitalizáciu vlastných vybraných fondov už od roku 2004 so zapojením na európske projekty financované z fondov programu Európskej komisie (projekt eTEN DOD 2006 – 2008 a EOD Culture 2009 – 2014). Fondy UKB neboli predmetom digitalizácie v rámci OPIS PO2.

Za viac ako 12 rokov svojej existencie si (dnes už samostatný) Odbor digitalizácie prešiel viacerými postupnými inováciami v oblasti metód skenovania, úprav digitalizačných pracovných postupov až po viaceré systémy digitálnych knižníc. So sprístupňovaním digitalizovaného knižného obsahu sme začali v roku 2007, kedy sme sprevádzkovali našu prvú digitálnu knižnicu s využitím voľne dostupného softvéru Greenstone, ktorú sme v roku 2011 nahradili systémom MediaINFO, ktorý používame dodnes.

Realizácia interných a medzinárodných projektov

V rámci bežnej prevádzky sme realizovali niekoľko interných digitalizačných projektov, počas ktorých sme komplexne zabezpečili čistenie, reštaurovanie, digitalizáciu, postprocessing a plnotextové sprístupnenie historicky významných bratislavských periodík v unikátnom vlastníctve Univerzitnej knižnice v Bratislave, napríklad Nyugatmagyarországi hiradó: politikai napilap a Pressburger Zeitung (v spolupráci s Maďarskou národnou knižnicou a DIFMOE). Pri významných digitalizačných projektoch nezabúdame na pasportizáciu všetkých ročníkov a jednotlivých čísiel a spolupracujeme s ďalšími pamäťovými inštitúciami pri výmene chýbajúcich čísiel a poškodených strán.

Digitalizačné služby

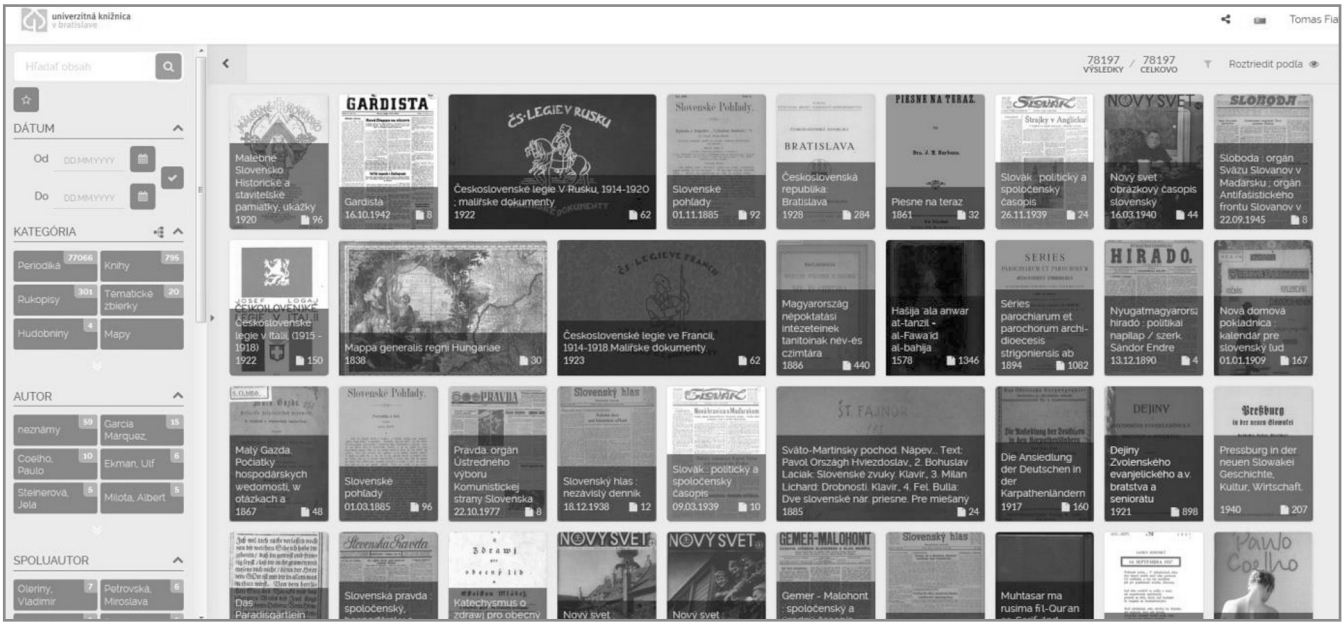
Odbor digitalizácie poskytuje široké spektrum služieb pre svojich používateľov, začínajúc skenovaním a elektronickým doručovaním častí dokumentov z mikrofilmov, kníh a periodík, pokračujúc zabezpečovaním skenovania v rámci požiadaviek národnej a medzinárodnej medzikniničnej výpožičnej služby, digitalizáciou voľne dostupných kníh v rámci európskeho konzorcia E-book on Demand (EOD), končiac prevádzkou samoobslužných skenerov pre používateľov. Jedným z posledných trendov v oblasti asistovaných digitalizačných služieb je každoročné znižovanie záujmu zo strany používateľov a preferencia samoobslužného skenovania. Pokles záujmu o asistované digitalizačné služby zrejme súvisí aj so zvýšením voľne dostupného objemu digitalizovaných dokumentov na internete, vyššou technologickou gramotnosťou používateľov našej knižnice a preferenciou služieb, ktoré sú dostupné zdarma.

Trendy v oblasti sprístupňovania

Z hľadiska efektívnosti sprístupňovania digitalizovaného obsahu je nutné zabezpečiť prístup k digitálnej knižnici naprieč všetkými technickými zariadeniami (smartfóny, tablety a stolné počítače) bez akýchkoľvek obmedzení, to znamená zabezpečenie vysokej responzívnosti webu a všetkých jeho funkcionalít. Medzi ďalšie súčasné trendy môžeme zaradiť sprístupňovanie čo najväčšieho počtu autorsky nechránených dokumentov voľne (mimo priestorov inštitúcie), rozšírenú prácu s plným textom, vytváranie používateľských zbierok, zdieľanie obsahu (sociálne siete a hyperlinky) konkrétnej časti zdigitalizovanej strany, oprava automatického optického rozpoznania znakov používateľmi systému, automatické prepájanie obsahu na informačné systémy, automatické strojové preklady plných textov, používateľské vyselektovanie zón záujmu z konkrétnej strany (fotografie, odseky a zóny).

Digitálna knižnica UKB

Prostredníctvom digitálnej knižnice (systém MediaInfo) plnotextovo sprístupňujeme viac ako 1100 000 strán digitalizovaných slovacikálnych dokumentov, kníh a periodík z fondu UKB. Kompletný zoznam sprístupnených dokumentov v digitálnej knižnici nájdete na webovej stránke UKB. Naša digitálna knižnica *http://digitalna.kniznica.info* umožňuje extrakciu textu, extrakciu ľubovoľnej zóny, extrakciu obrazu, preklad textu, metadátový popis vo viacerých jazykoch, vyhľadávanie podľa slov na úrovni celej zbierky, ale i konkrétneho dokumentu, zoskupovanie obsahu podľa faziet v kombináciách s inými kritériami (dátum vydania, výskyt frázy v obsahu). Z hľadiska funkcií pre zaregistrovaných používateľov sú významné aj funkcie poznámok v texte, označenie obľúbeného obsahu a možnosti ich zdieľania s ostatnými používateľmi. Digitálna knižnica UKB je multifunkčným nástrojom, ktorý vie efektívne sprístupňovať širokú škálu digitálneho obsahu (audio, video, dáta) s možnosťou funkcionality inštitucionálneho repozitára.



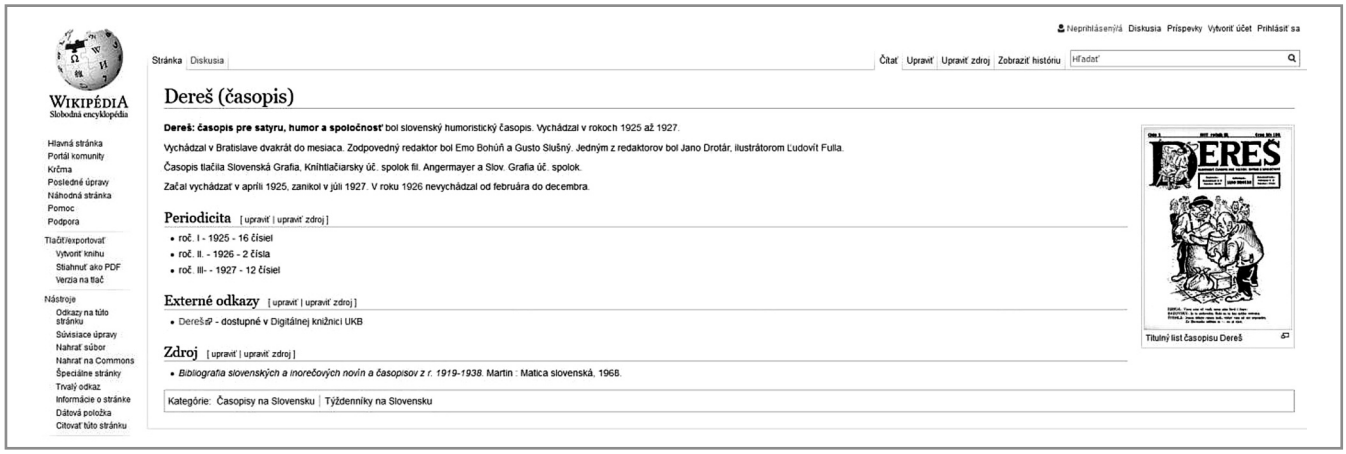
Obrázok č. 2 Titulná stránka Digitálnej knižnice UKB (generovaná dynamicky na základe najnavštevovanejších dokumentov).

Indexovanie, viditeľnosť a integrácia digitálnych objektov na internet

Predpokladom pre efektívne nájdenie digitalizovaných knižničných dokumentov na internete je sprístupnenie a zindexovanie bibliografických údajov vyhľadávacími robotmi, čo zabezpečí zvýšenú viditeľnosť digitálnej knižnice. Podobne je možné sprístupniť na indexáciu plnotextový index slov vytvorený na základe automatického optického rozpoznania digitalizovaných knižných dokumentov. Plnotextový index predstavuje obrovské množstvo dát (v UKB cez miliardu slov), ktoré je technicky možné z veľmi malej časti sprístupniť a nechať zindexovať vyhľadávacími robotmi. Po realizácií vyššie spomínaných krokov očakávame zvýšenú návštevnosť a nájditeľnosť plnotextového digitalizovaného obsahu. Medzi ďalšie možnosti patrí manuálne editovanie relevantných zdrojov pre vyhľadávače, napríklad vytváranie a obohacovanie článkov vo Wikipédii. Podstatné je aj informovanie a zverejňovanie informácií o najnovšie sprístupnených dokumentoch v digitálnej knižnici, prostredníctvom dynamicky generovaného obsahu odosielaného priamo do koncových redakčných systémov. V Univerzitnej knižnici sa snažíme zvýšiť a zviditeľniť svoje digitalizované dokumenty predovšetkým dopĺňaním externých liniek na zdigitalizovaný obsah na Wikipédiu, pri neexistujúcich stránkach na Wikipédii zabezpečujeme ich vytvorenie. Zároveň pracujeme na sprístupnení bibliografických údajov a častí plných textov pre robotov internetových vyhľadávačov.

Sprístupňovanie digitalizovaných dokumentov – copyright

Digitalizované dokumenty sprístupňujeme v zmysle platnej legislatívy v rámci priestorov knižnice voľne, mimo knižnice len autorsky nechránené dokumenty, ktorých autor je 70 a viac rokov po smrti.



Obrázok č. 3 Ukážka vytvorenej stránky časopisu Dereš vo Wikipédii s linkou do Digitálnej knižnice.

Sme presvedčení, že súčasné znenie autorského zákona umožňuje prekročiť túto métu, radi by sme prostredníctvom vzdialeného prístupu začali sprístupňovať aj autorsky chránený obsah našim čitateľom. Zabezpečenie verifikácie používateľa, zamedzenie sťahovania dokumentu a vykazovanie prístupov k jednotlivým dokumentom pre účely finančnej odplaty autorským zväzom nepredstavujú pre nás technický problém. Momentálne čakáme na hlbšiu právnu analýzu, ktorá by mohla dať projektu zelenú.

Recyklovanie digitálnych objektov a vytváranie nových informačných produktov

Rôznorodý obsah digitálnej knižnice (články, strany, odseky) z rôznych knižničných dokumentov (periodiká, knihy, mapy) je možné na základe rôznych asociácií, vlastností, udalostí a faktov pospájať do nového komplexného recyklovaného informačného produktu, ktorý agreguje informácie z viacerých zdrojov a vytvára komplexný prehľad entít (osôb, miest, udalostí), súvislostí a faktov v graficky znázornených väzbách, pojmových mapách, ontológiách alebo v hypertexte. Digitálna knižnica UKB už dnes disponuje funkcionalitami integrovaného autonómneho redakčného systému na recyklovanie importovaných digitálnych objektov a vytváranie jednoduchých informačných prehľadov. V budúcnosti plánujeme funkcionalitu rozšíriť o automatizované vytváranie prehľadových stránok na základe spájania rôznych entít s prepojením na znalostné bázy a vizualizáciou vzťahov.

Automatické prepájanie na metadátové informačné zdroje

Medzi rozšírené funkcionality digitálnych knižníc patria aj možnosti prepojenia na knižničný katalóg so zabezpečením transferu existujúcich bibliografických metadát, ich obohatením a integráciou s digitalizovanými objektmi. Oba systémy by mali byť schopné medzi sebou komunikovať a vymieňať si metadátové informácie. Aktuálne riešime možnosť automatického preberania bibliografických záznamov zo súborného katalógu periodík a knižničného katalógu UKB do systému digitálnej knižnice so súbežnou editáciou pôvodných bibliografických záznamov s doplnením hyperlinky na zdigitalizovaný obsah v poli 856 (Marc 21).

Reštaurovanie a digitalizácia

Reštaurátorský zásah počas digitalizačného procesu je neodmysliteľnou súčasťou pri digitalizácii historických knižných dokumentov (najmä periodík), pričom pomáha uvoľnením viazanej väzby (čím zabezpečí lepšie podmienky pre digitalizáciu a vyššiu kvalitu skenov). Pred digitalizáciou vyčistíme strany od prachu a nečistôt, natrhnuté časti vylepíme japonským papierom, pokrčené listy jemne navlhčíme a lisujeme do listinnej podoby. Po digitalizácii dokumenty viažeme späť do pôvodných dosiek, pričom spôsob väzby jednotlivých strán vyberáme podľa spôsobu ďalšieho používania dokumentu. Týmto prístupom zabezpečíme neobmedzený prístup k naskenovaným dokumentom, fyzické originály zreštaurujeme a zároveň ich správnym uložením a obmedzením ich používania zamedzíme ďalšej degradácii papiera. (obrázok č. 4)

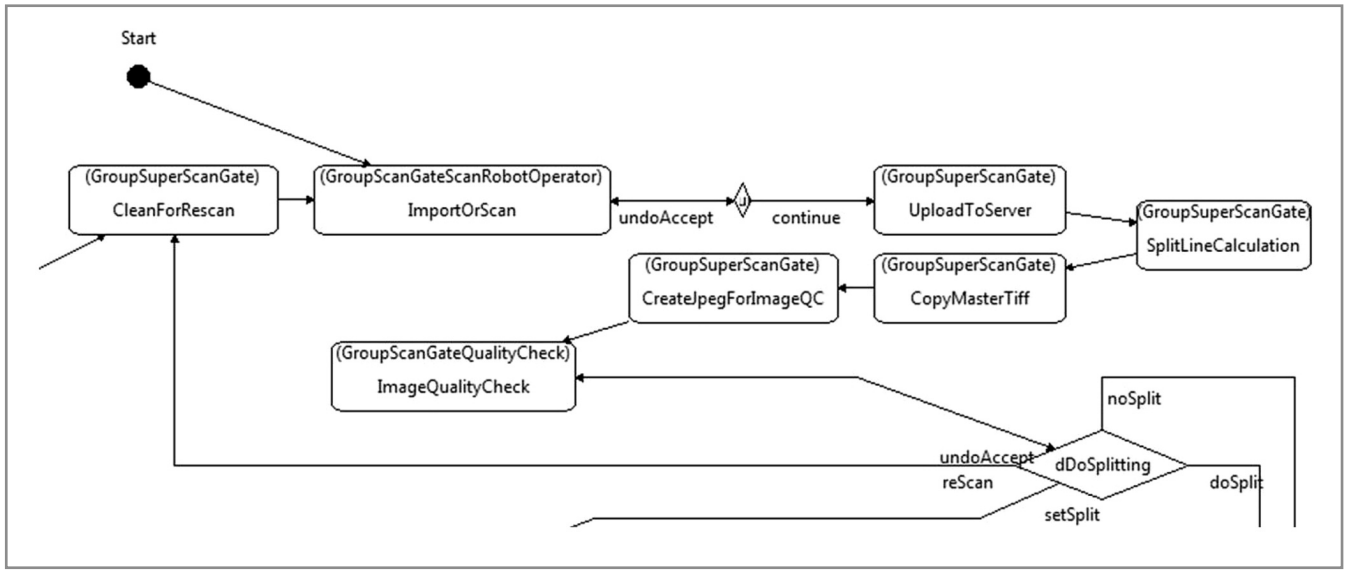
Digitalizačný workflow a automatizácia

Väčšina procesov v Odbore digitalizácie je plne automatizovaná. V rámci riadenia interných procesov používame softvér Scanflow od spoločnosti Treventus. Medzi základné funkcionality softvéru patrí delegovanie pracovných úloh jednotlivým používateľom, prideľovanie špecifických pracovných úloh konkrétnym používateľom, kontrola kvality, štatistické vyhodnotenia úloh a používateľov, ako aj absolútny prehľad o stave jednotlivých pracovných úloh. Automatizovaný workflow má prepojenie na ABBYY Recognition server a na interné úložisko, kam sa ukladajú zálohy jednotlivých pracovných úloh, výstupné formáty, ale i pôvodné skeny. Všetky súbory, adresáre, majú v sebe zakomponovaný jedinečný identifikátor, prostredníctvom ktorého vieme jasne identifikovať digitalizovaný objekt až na úroveň dátumu. (obrázok č. 5)

Optické rozpoznávanie znakov

Pre zabezpečenie plnotextového vyhľadávania v digitálnej knižnici používame dve inštalácie ABBYY Recognition server. Jeden zo serverov so špeciálnou licenciou na počet strán je určený na rozpoznávanie švabachu, druhý server slúži na spracovanie

Obrázok č. 4 Stav dokumentu pred a po reštaurovaní.



Obrázok č. 5 Ukážka niektorých procesov v Digitalizačnom workflow.

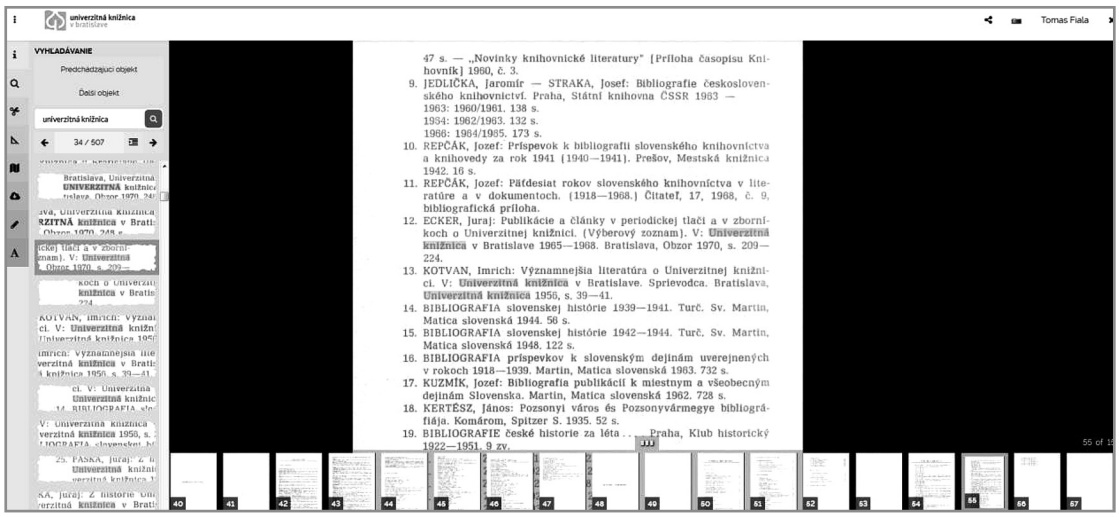
súčasných typov písma. Výstupnými formátmi z optického rozpoznávania znakov je plnotextovo prehľadateľné PDF (v MRC kompresii) a ALTO xml s koordinátmi jednotlivých znakov (používa sa na presné spárovanie obrazu a plného textu v digitálnej knižnici).

Formáty

Primárnym formátom, ktorý náš odbor používa je TIFF, ostatné formáty (PDF, JPEG, XML, TXT, HTML, HTM, JP2 slúžia len ako deriváty pre špecializované aplikácie. Z hľadiska prirodzeného vývoja formátov sme počas dvanásťročnej prevádzky pracoviska zaznamenali zmeny vo verziách a štandardoch. Z formátov, ktoré používame sa najdynamickejšie vyvíjal formát PDF, z ktorého máme identifikovaných päť verzií. Za ním nasleduje formát JPEG z ktorého máme identifikované dve verzie. Pre najjednoduchšie zabezpečenie dlhodobej archivácie (LTP) je vhodné používať čo najmenej formátov, situáciu komplikujú ich časté zmeny. Samotná identifikácia formátov a ich verzií nad veľkým objemom dát, ktorý sa počas rokov vygeneroval, predstavuje problém, nehovoriac o zvýšenej náročnosti na dlhodobú archiváciu.

Výskum a vývoj v oblasti automatického spracovania obsahu

Odbor digitalizácie plánuje v tomto roku spustiť projekt so Slovenskou technickou univerzitou, Fakultou informatiky a informačných technológií, s ktorou chceme spoločne riešiť vedeckovýskumný projekt identifikácie entít a ich vzťahov nad opticky rozpoznanými výstupmi z procesu digitalizácie knižničných dokumentov s grafickou vizualizáciou vzájomných väzieb. V rámci projektu sa plánuje spracovanie digitalizovaných objektov (jednotlivých čísiel novín) s cieľom automatizácie procesu analytického rozpisu článkov. Prvým predpokladom je naprogramovanie softvéru, ktorý bude automaticky rozpoznávať jednotlivé



Obrázok č. 6 Ukážka plnotextového vyhľadávania v Digitálnej knižnici.

články s identifikáciou, extrakciou metadát a tvorbou článkových bibliografických záznamov. V ďalších častiach projektu by sme radi rozpoznali základné entity z textov článkov a prepojili ich s inými znalostnými databázami, čím by sme vytvorili nástroj na vizualizáciu znalostnej databázy a zdrojov. V poslednej časti projektu plánujeme vytvoriť základný nástoj pre tvorbu virtuálnych výstav a prezentácií.

Obrázok č. 7 Ukážka grafickej vizualizácie slovných väzieb na základe analýzy plných textov naskenovaných dokumentov v Digitálnej knižnici.

Mgr. Tomáš Fiala
tomas.fiala@ulib.sk
(Univerzitná knižnica v Bratislave)