Digitálne pramene – webharvesting a archivácia e-Born obsahu

Alojz Androvič
Alojz.androvic@ulib.sk

Ivan Ciglan
ivan.ciglan@ulib.sk

Jana Matúšková
jana.matuskova@ulib.sk

V apríli 2015 bola Univerzitná knižnica v Bratislave poverená riešením národného projektu Digitálne pramene – webharvesting a archivácia e-Born obsahu. Projekt bol realizovaný v rámci Operačného programu Informatizácia spoločnosti (OPIS) a spolufinancovaný Európskym fondom regionálneho rozvoja. Informačný systém Digitálne pramene na zber, identifikáciu, manažment a dlhodobú ochranu webových prameňov a e-Born dokumentov pozostáva zo špecializovaných, prevažne voľne dostupných softvérových modulov. Aplikáciu podporuje výkonná hardvérová infraštruktúra. Celková kapacita archívu IS Digitálne pramene je 800 TB. Centrálny dátový archív bude slúžiť ako úložisko pre dlhodobé uchovávanie. Projekt bol ukončený k 31. 12. 2015. Počas pilotnej prevádzky sa uskutočnili tri typy zberov: komplexný, tematický a výberový. V súčasnosti sa realizuje rutinná prevádzka, ktorú zabezpečuje oddelenie Depozitu digitálnych prameňov so špecializovanými digitálnymi kurátormi. Prax a realizácia projektu sú výrazne ovplyvnené aktuálne platnou legislatívou.

Úvod

Znie to už priam ako klišé, že internet mnoho zmenil, no faktom zostáva, že aj po mnohých rokoch sme si ešte nie celkom uvedomili všetky aspekty jeho existencie. Špecifiká elektronického publikovania a z neho vyplývajúce rôzne prístupy k publikovaniu informácií, ich volatilita (nestálosť), živelnosť, (ne)dôveryhodnosť a dynamika nás nútia prehodnocovať prístupy k hodnoteniu tohto fenoménu. Po stáročiach sa vyvinul nový druh publikovania a sprístupnenia informácií a ľudstvo bolo nútené v priebehu pár rokov zmeniť zaužívané vzorce, ktorými manipuluje s publikovaným obsahom.

Trvalo roky, kým sme pochopili, že obsah, ktorý prináša internet, je bez ohľadu na jeho kvalitatívne aspekty hodný uchovania pre ďalšie generácie. Táto myšlienka spustila po celom svete iniciatívy na ochranu pôvodného digitálneho obsahu a jeho zaradenie ako súčasti národného kultúrneho dedičstva, ktoré vzniká bez fixácie na materiálny nosič. Jednou z týchto iniciatív je aj projekt „Digitálne pramene – webharvesting a archivácia e-Born obsahu", ktorý sa podarilo na Slovensku realizovať s podporou finančných prostriedkov z fondov EÚ v celkovej výške 6,8 milióna EUR v rámci Operačného programu Informatizácia spoločnosti – Prioritná os 2 (OPIS PO2). Vo sfére kultúry dostala Univerzitná knižnica v Bratislave (UKB) jedinečnú príležitosť byť dôležitým garantom, ktorý bude chrániť webový obsah a pôvodné elektronické „e-Born" dokumenty pred ich nenávratnou stratou.

Problematika zberu, spracovania, archivácie a následného využitia pôvodného digitálneho obsahu je veľmi dynamická a rôznorodá. Slovensko vstúpilo projektom do tejto iniciatívy pomerne neskoro. Na jednej strane nás to pripravilo o množstvo nenávratne strateného obsahu za posledné roky, na strane druhej nám poskytlo výhodu porovnať si existujúce riešenia a prístupy a na základe ich evaluácie vybrať vhodné metodické a technicko-technologické riešenie. V decembri 2014, keď sa projekt dostal do aktualizovaného zoznamu projektov OPIS PO2, začala predprojektová príprava a podnikli sa kroky pre vytvorenie podmienok na jeho realizáciu. Projektom bola poverená Univerzitná knižnica v Bratislave (UKB), ktorej sa podarilo zúročiť skúsenosti získané pri realizácii projektu „Centrálny dátový archív" a následne vytvoriť komplexné riešenie pre záznam, zbieranie, sprístupňovanie a ochranu pôvodného digitálneho obsahu, webharvesting a webarchiving online digitálnych prameňov ako integrálnej súčasti kultúrneho dedičstva Slovenskej republiky.

Realizácia projektu

Riešenie projektu začalo oficiálne 1. 4. 2015 s plánovanou dobou realizácie 9 mesiacov.

V mesiacoch august až december 2015 sa realizovala implementácia a pilotná prevádzka projektu. Vybudovala sa rozsiahla hardvérová a softvérová infraštruktúra na zber, spracovanie, archiváciu a indexovanie elektronického obsahu. Informačné a komunikačné technológie (IKT) boli nainštalované v priestoroch dátového centra „Centrálneho dátového archívu" na Klariskej ulici. Následne prebehlo testovanie a odlaďovanie aplikačných a hardvérových komponentov. V decembri 2015 boli vymenované dve komisie na preberanie implementácie IKT infraštruktúry a Informačného systému Digitálne pramene, ktoré formou akceptačných a užívateľských testov overili splnenie funkčných a nefunkčných požiadaviek podľa schválenej projektovej dokumentácie.

Počas druhého polroka prebiehali na pravidelnej báze rokovania a pracovné stretnutia s dodávateľom ohľadom hardvérového a softvérového riešenia, metodických otázok prípravy systému na správu a uchovávanie e-Born obsahu, katalógu webových stránok a repozitára e-Born dokumentov. Bolo potrebné vyriešiť a stabilizovať problematiku metadátových modelov pre webové stránky i e-Born dokumenty, pripravovala sa verejná stránka a portál, prostredníctvom ktorého sa sprístupňujú údaje o e-Born dokumentoch. Pôvodná webová stránka pre projekt Digitálne pramene (www.webdepozit.sk) bola vytvorená v redakčnom systéme Drupal. Začiatkom roku 2016 ju nahradila nová a optimalizovaná verzia stránky vo WordPress-e, ktorý poskytuje užívateľsky prívetivejšiu administráciu a rozšírenú funkcionalitu.

Obr. 1 Digitálne pramene – webové sídlo www.webdepozit.sk

V rámci realizácie projektu sa pracovalo aj na príprave metodických materiálov. Boli vypracované rámcové dokumenty ako „Politika zberu webových prameňov" a „Politika zberu e-Born dokumentov" (zvlášť pre seriály a pre monografie), ako aj metadátové modely pre tieto typy prameňov. Súčasťou metodických materiálov sú tiež dokumenty: „Štatút DIP (Digitálne pramene)" a „Určené spoločenstvo DIP."

Pilotná prevádzka projektu Digitálne pramene – webharvesting a archivácia e-Born obsahu prebiehala v mesiacoch november až december 2015. V rámci nej sa uskutočnil pilotný celodoménový (komplexný) zber aktívnych domén vo webovom priestore.sk. Zber prameňov vychádzal zo zoznamu slovenských domén od spoločnosti SK-NIC, správcu národnej domény .sk. Ostatné domény ako napr., .com, .org, .net, .eu boli doplnené v prípade, že spĺňali kritérium slovacikálneho charakteru. Z hľadiska charakteru ich evidencie je tieto domény nutné vyhľadávať manuálne.

Výsledky pilotného celodoménového zberu:

Počas pilotnej fázy projektu sa uskutočnil tiež tematický zber webových stránok 3000 subjektov, ktoré sú súčasťou informačného systému „Kultúrny profil Slovenska." Boli to predovšetkým webové stránky z oblasti kultúry, školstva, štátnej a verejnej správy a z tretieho sektora.

Počet záznamov pre zber:	332896
Počet úspešne zozbieraných domén:	241717
Objem zozbieraného obsahu (spolu)	48 TB
Počet zozbieraných objektov (spolu)	997 mil.

Tretím typom zberu v rámci pilotnej fázy projektu bol výberový zber 10 elektronických online seriálov, ktoré boli vybrané v spolupráci s Národnou agentúrou ISSN. Pilotná prevádzka projektu bola k 31. decembru 2015 ukončená. Od januára 2016 pokračuje projekt v rutinnej prevádzke [1].

Architektúra systému

Riešenie softvérového subsystému DIP je navrhnuté modulárne s jednotným GUI front-end-om. Pozostáva celkovo z nasledovných 13 logických modulov postavených na open-source riešeniach:

Databáza domén – obsahuje zoznam všetkých sledovaných domén (nielen z .sk priestoru) a základné atribúty o nich. Je východiskom pre plánovanie výberového a tematického zberu.

Databáza domén je chápaná ako headless komponent, s ktorým používatelia pracujú pomocou modulu Kurátor. Táto databáza je využívaná technickými procesmi.

Úložisko archivovaných WARC súborov a newebového obsahu – zdieľaný súborový systém, na ktorý sa ukladajú súbory WARC vytvorené Harvesterom.

Harvester – slúži na zber a archiváciu vybraných webových stránok v podobe súborov WARC. Súčasťou je aj riešenie (add-on) na deduplikáciu ukladaného obsahu.

Za logickú súčasť Harvestera sa považuje aj ukladanie newebového obsahu vkladaného cez Portál Pôvodcom newebového obsahu.

Prehliadač uloženého obsahu deduplikáciu slúži na zobrazenie obsahu uloženého vo WARC súboroch.

Metadátový index – v procese harvestingu sa vytvára aj metadátový index pre potreby následného vyhľadania. Metadáta sa extrahujú hlavne z html/head sekcie a podľa potreby sa kombinujú s metaúdajmi z doménovej databázy. Tento modul je chápaný ako headless komponent, s ktorým používatelia pracujú pomocou modulu Portál.

Portál – má verejnú a privátnu časť. Vo verejnej časti vytváranej pomocou CMS sú publikované a udržiavané relevantné informácie o projekte. Verejná časť obsahuje aj GUI na vyhľadanie metódou full-text alebo kombinácie metadát. Privátna časť slúži okrem iného aj autorizovaným Pôvodcom newebového obsahu na upload newebového obsahu.

Integrácia s CDA – tento funkčný celok zabezpečuje zabalenie WARC a metadát do SIP balíkov a ich vklad do CDA. Predpokladá sa online vklad s prípadnou explicitnou žiadosťou o vklad.

Integrácia s SK-NIC – modul zabezpečuje periodické načítanie zoznamu slovenských domén z www.sk-nic.sk, zistenie ich reálneho funkčného stavu a aktualizáciu ich záznamu v Databáze domén.

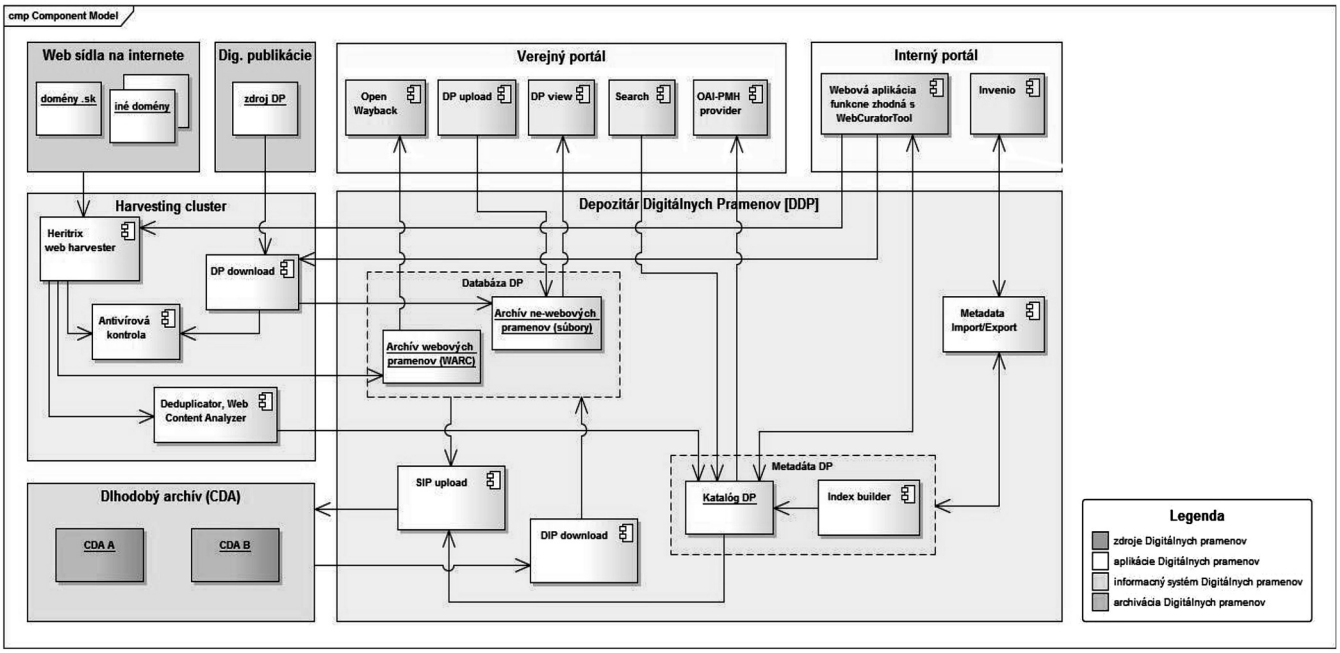
Kurátor – modul slúži na prácu nad Databázou domén a na plánovanie zberu webového obsahu. Prístup je pomocou GUI, vykonané zmeny sa zapisujú do Databázy domén a riadiacich súborov Harvestera.

Invenio – modul obsahuje alternatívnu evidenciu záznamov o doménach a uchovávanom newebovom obsahu, a to vo forme opisných metaúdajov v štandarde MARC 21. V module je možné vykonávať štandardné knižničné operácie nad uloženými záznamami.

OAI-PMH – modul vystavuje metadáta z Databázy domén pre potreby ich zberu tretími stranami.

Auditing – modul centralizuje spracovanie audit záznamov.

Infraštruktúrne aplikácie do tohto logického komponentu tvoria rôzne podporné infraštruktúrne aplikácie typu Monitoring, Antivirus a pod.



Obr. 2 Diagram komponentov IS Digitálne pramene

IS Digitálne pramene je postavený na otvorených softvérových riešeniach ako Linux, Apache, Apache Tomcat, PostgreSQL, Heritrix, DeDuplicator, OpenWayback, Invenio, SOLR a pod. Integrácia komponentov do homogénneho funkčného celku je zabezpečená vlastným vývojom aplikačného rozhrania na báze programovacieho jazyka Java. Aplikačná vrstva beží na systémovej platforme RedHat Linux.

Infraštruktúra

Predpokladom na úspešné zvládnutie procesov zberu a spracovania veľkých dátových objemov, ktoré sú typickou črtou webharvestingu, je primerane výkonná technologická infraštruktúra.

Navrhované riešenie (Obr. 3) je optimalizované na paralelné spracovanie, harvestovanie a indexáciu archivovaného obsahu. Virtualizačnú platformu tvorí skupina fyzických serverov, ktorá pomocou hypervízora serverovej virtualizácie poskytuje virtuálne servery pre aplikačné moduly. Virtualizačná platforma je inštalovaná na 21 fyzických serveroch. Každý z týchto serverov poskytuje

- 2x CPU 2,3 Ghz a 12 corov
- 256 Gb RAM
- 2x SSD Disk 200 GB
- 2x 10 Gbit LAN, 2x 8 Gbit Fibre channel

Na tejto platforme sa aktivuje podľa potreby skupina virtuálnych serverov, workerov. Je to základný modul webharvestingu. Realizuje spracovanie konkrétnej stránky alebo podstránky a vykoná všetky požadované operácie. Výsledkom sú spracované údaje v požadovanej forme. Worker obsahuje všetok softvér potrebný k transformácii webovej stránky na archívny balík vo formáte WARC. Úlohy mu prideľuje Director, ktorý riadi a vykonáva dohľad nad spracovávaním úloh. Workery využívajú spoločný zdieľaný diskový priestor.

Tri fyzické servery slúžia ako cluster databáz potrebných pre aplikácie a zároveň zobrazujú zdieľané údaje pre potreby ostatných serverov. Tieto servery nie sú virtualizované, lebo existuje predpoklad ich plného fyzického využitia pre potreby databáz a diskového úložiska. Skupina databázových serverov slúži na ukladanie archívnych balíkov a metaúdajov ku harvestovaným dátam. Medzi jednotlivými databázovými uzlami beží synchrónna (multi-master) replikácia.

Pre potreby spracovania, indexácie a archivácie je systém DIP vybavený 800 TB diskovým subsystémom, ktorý spĺňa požiadavky pre zabezpečenie adekvátneho pracovného a úložného priestoru pri zachovaní potrebnej redundancie a výkonových parametrov. Migrácia dát medzi jednotlivými zónami prebieha automaticky a je rozdelená do troch úrovní reprezentovaných „rýchlymi" a „pomalými" diskami. Diskový susbsystém tvoria SSD, SAS a SATA disky rozdelené analogicky v percentuálnom pomere 3 : 30 : 67. Diskové polia slúžia ako primárne blokové dátové úložiská. Sú pripojené cez SAN (Storage Area Network) ku serverom cez 8 ciest, každá z nich má kapacitu 8 Gbps. Diskové polia majú medzi sebou synchrónnu replikáciu dát, replikujúcu dáta na úrovni blokov. Tento prístup zaručuje dostupnosť dát v prípade výpadku jedného diskového poľa.

Z dôvodu veľkých prenášaných dátových objemov sa na prepojenie infraštruktúrnych komponentov (serverov, diskov, prepínačov, ...) používajú vysokorýchlostné rozhrania Ethernet (10 Gbit) a FibreChannel (8 Gbit). Konektivita navonok je zabezpečená prostredníctvom siete SANET (1 Gbit) a je chránená firewallmi. Firewall (FW-Internet) oddeľuje internet od demilitarizovaných zón, ktoré obsahujú služby dostupné z internetu a služby komunikujúce s externými systémami, napríklad s Centrálnym dátovým archívom (CDA). Firewall (FW-VPN) oddeľuje internet od administrátorských služieb (Intra portál.)

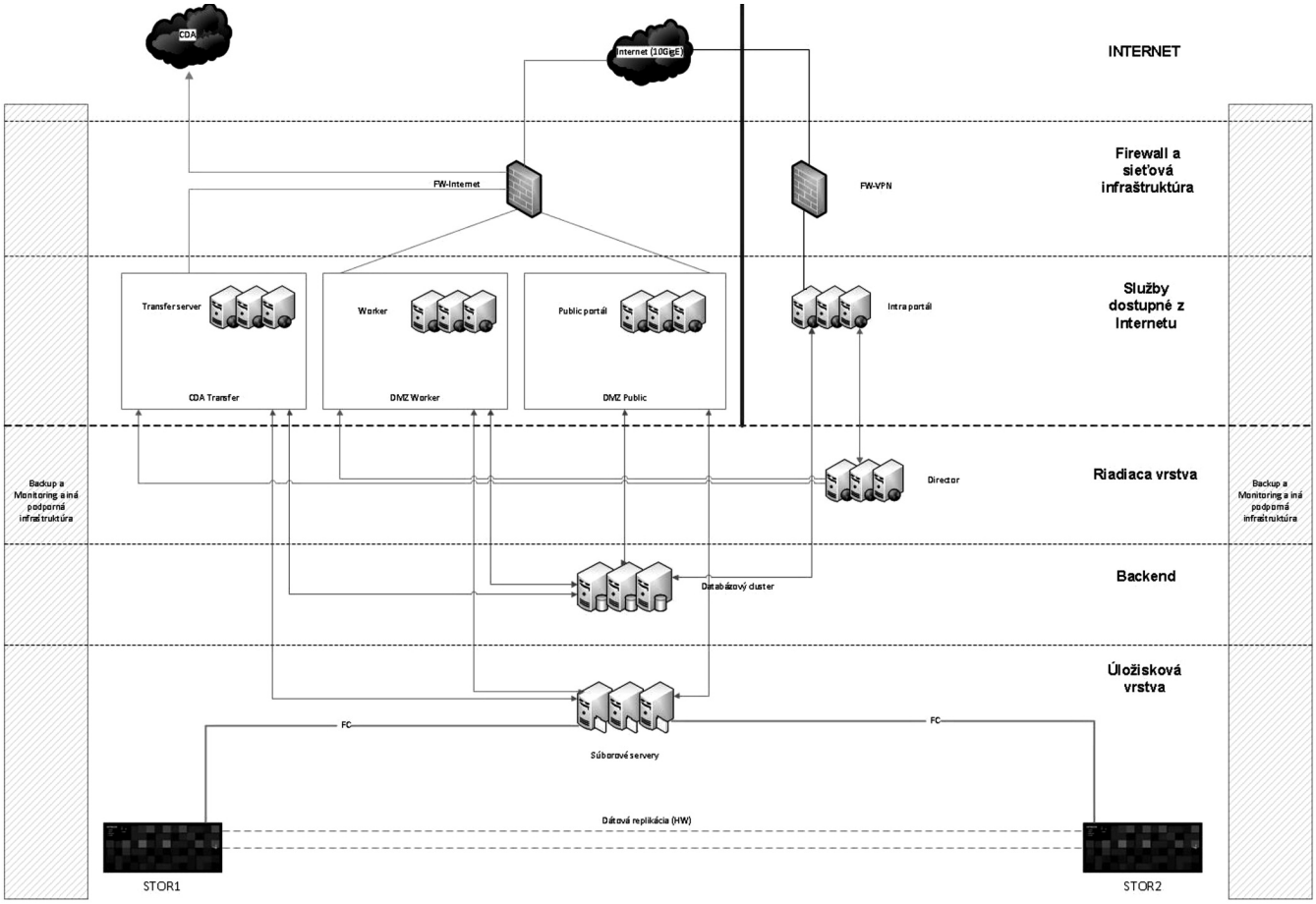
Od procesu spracovania je oddelená fyzická vysoko dostupná infraštruktúra, ktorá poskytuje obslužné infraštruktúrne služby. Zálohovací systém (backup) slúži na zálohovanie systémových prostriedkov infraštruktúry. Zálohujú sa operačné systémy a aplikácie a ich konfigurácie. Zálohovacím médiom je pásková knižnica. Harvestované údaje z internetu budú zálohované po ich spracovaní v rámci infraštruktúry transferom do CDA. Monitoring využíva systém Zabbix, ktorý sleduje a monitoruje výkonové a chybové stavy komponentov infraštruktúry a notifikuje ich. Zároveň sleduje základné parametre operačných systémov a aplikácií.

Organizačné zabezpečenie

Na zabezpečenie úloh súvisiacich s prípravou, riadením, implementáciou a prevádzkou národného projektu Digitálne pramene – webharvesting a archivácia e-Born obsahu bolo v rámci organizačnej štruktúry Univerzitnej knižnice v Bratislave vytvorené samostatné oddelenie Depozit digitálnych prameňov (DDP). Oddelenie vzniklo dňa 1. 6. 2015 a pôsobí v rámci odboru Národná agentúra ISSN a Depozit digitálnych prameňov v štruktúre Úseku elektronizácie a integrácie UKB. Oddelenie má štyroch pracovníkov – troch digitálnych kurátorov a jedného vedúceho.

Digitálne kurátorstvo a digitálny kurátor

Digitálne kurátorstvo je odbor, ktorý rieši dlhodobú ochranu digitálnych informácií pomocou postupov založených na udržiavaní a zvyšovaní hodnoty uchovávaných dát. Ako uvádza vo svojej diplomovej práci Michal Konečný, „nestačí mať



Obr. 3 Infraštruktúra IS Digitálne pramene

dáta bezpečne uložené, je potrebné sa o ne starať rovnako starostlivo a odborne, ako sa kurátori starajú o exponáty v múzeách a galériách" [2]. Problematike digitálneho kurátorstva sa v slovenských podmienkach podľa našich informácií zatiaľ nikto nevenoval. Historický prehľad vývoja definícií termínu digitálne kurátorstvo prináša vyššie spomenutá diplomová práca. Pre naše účely z hľadiska činností uskutočňovaných kurátormi v Depozite digitálnych prameňov sa najviac hodí definícia na stránkach Digital Curation Centre (vzniklo v r. 2004 ako medzinárodné centrum pre vývoj nástrojov a techník pre dlhodobé a bezpečné dátové kurátorstvo): „Digitálne kurátorstvo zahŕňa správu, ochranu a zhodnotenie digitálnych výskumných dát počas ich životného cyklu [3]."

Termín digitálne kurátorstvo je úzko spätý s termínmi digitálne uchovávanie (digital preservation) a dlhodobé uchovávanie (long term preservation). Podľa definície Neila Beagrieho je digitálne uchovávanie „rad riadených aktivít nevyhnutných k zabezpečeniu kontinuálneho prístupu k digitálnym materiálom [4]." Dlhodobé uchovávanie zahŕňa súbor aktivít a postupov, ktorých cieľom je zabezpečiť použiteľnosť, vyhľadateľnosť, dostupnosť a autenticitu digitálneho obsahu v priebehu času. Digitálne kurátorstvo tvorí akúsi nadstavbu dlhodobého uchovávania, nakoľko sa zameriava nielen na dlhodobé uchovávanie dát, ale aj na dlhodobú ochranu ich užitočnosti a použiteľnosti [2]. Ako ďalej uvádza M. Konečný, v súčasnosti už vo svete predstavuje digitálne kurátorstvo etablovaný odbor s vlastnou históriou, inštitúciami a odbornou komunitou. Na viacerých zahraničných univerzitách je možné tento odbor študovať formou rôznych kurzov, špecializovaných predmetov či v rámci doktorandského štúdia (The University of California Curation Center, The Digital Curation Institute na University of Toronto). Hlavnými témami sú: plány a stratégie dlhodobého uchovávania, formáty a formátové politiky, metadáta, OAIS, výber a hodnotenie, metódy uchovávania, dôveryhodný digitálny archív, sprístupnenie dát a ďalšie.

V praktickej časti spomínanej diplomovej práce navrhuje autor kompetenčný model digitálneho kurátorstva a obsahovú náplň vysokoškolského magisterského kurzu pre ČR pod názvom Úvod do digitálneho kurátorstva. Kompetenčný model digitálneho kurátorstva zahŕňa podľa Konečného 15 kompetencií, ktoré by mal zvládnuť digitálny kurátor. Ide o tieto kompetencie: znalosť problematiky dlhodobého uchovávania, porozumenie potrebám určenej skupiny používateľov, výber a hodnotenie obsahu, príjem a správa dát, znalosť formátov a dátových štruktúr, prevod formátov, bezpečnosť a sprístupnenie, znalosť technologických aspektov dlhodobého uchovávania, sledovanie zmien formátov a technológií, zabezpečenie interoperability a dôveryhodnosti, analytická a metodická kompetencia, správa, riadenie a plánovanie digitálneho archívu, komunikácia a manažment, osveta a vzdelávanie.

Práca digitálnych kurátorov v našej praxi

Prácu digitálnych kurátorov Depozitu digitálnych prameňov v UKB riadi vedúci oddelenia. Je zodpovedný za chod oddelenia, usmerňuje a kontroluje činnosť kurátorov, má na starosti metodickú činnosť a vytváranie metodických postupov. Sleduje vývoj v oblasti legislatívy na Slovensku i vo svete, webarchivačné projekty v zahraničí a príklady – best practices. Participuje na výbere zdrojov pre archiváciu, vykonáva aktivity spojené s uzatváraním zmlúv o archivácii a sprístupňovaní elektronických on-line prameňov, komunikuje s poskytovateľmi a vydavateľmi a rieši problémy, ktoré vznikajú v súvislosti so zmluvami.

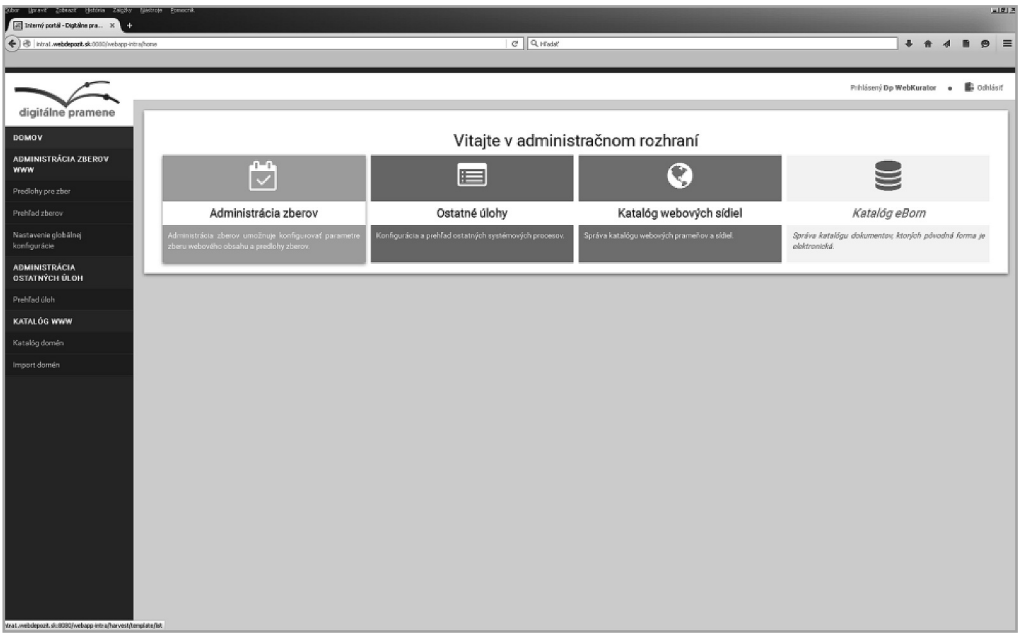
Kurátori sú špecializovaní (viď. nižšie), avšak navzájom úzko spolupracujú tak, aby sa dokázali vzájomne zastúpiť.

Kurátor I – analytik. Zabezpečuje správu a riadenie informačného systému, správu IKT zdrojov systému a databáz, rieši technické problémy, otázky bezpečnostnej politiky a interoperability. Stará sa o licencie a technickú dokumentáciu systému Digitálne pramene a v prípade potreby komunikuje s dodávateľom systému ohľadom technickej podpory. Aktívne spoluvytvára verejný portál webdepozit.sk.

Kurátor II – špecialista na webové pramene. Zabezpečuje návrh a realizáciu politiky zberu webových stránok, metodiku a identifikáciu webových domén a prameňov, komunikáciu s poskytovateľmi (oslovuje ich s návrhom na archiváciu, resp. preberá od nich návrhy na archiváciu webových sídiel), spravuje katalóg a archív webových stránok. Aktívne sa podieľa na budovaní a aktualizácii verejného portálu webdepozit.sk. Pred podpisom zmluvy o archivácii webových stránok ich testuje a kontroluje, ako sa zozbierajú. V prípade, že sa niektorá stránka nezozbiera správne, identifikuje problémy a zisťuje nastavenia robotov na stránke. V súlade s Politikou zberu webových stránok [5] sa pri zbere rešpektujú nastavenia robots.txt. Mnohé stránky, ktoré ich majú, nie je možné zozbierať v zodpovedajúcej kvalite. V týchto prípadoch kurátor komunikuje s poskytovateľmi webových stránok ohľadom ďalšieho postupu.

Kurátor III – špecialista na e-Born pramene. Zabezpečuje výber a akvizíciu e-Born obsahu, metodiku a identifikáciu komplexných digitálnych prameňov, správu katalógu a repozitára e-Born obsahu (e-seriály, e-monografie). Komunikuje s jednotlivými vydavateľmi, oslovuje ich s návrhom na archiváciu, resp. preberá návrhy na archiváciu e-Born prameňov, inštruuje vydavateľov o tom, ktoré metatagy a metadáta je treba vyplniť, aby boli ich publikácie kvalitne zaindexované a vyhľadateľné (systém podporuje fulltextové vyhľadávanie). Pred podpisom zmluvy dohodne kurátor spôsob dodávania, resp. ukladania e-Born obsahu do archívu. Dodávanie je možné dvomi spôsobmi – uploadom zo strany vydavateľa alebo downloadom zo strany kurátora.

K ďalším úlohám, na ktorých sa podieľajú všetci traja kurátori, patrí hodnotenie a testovanie webových stránok a e-Born prameňov z hľadiska kvality ich zberu, spúšťanie a vyhodnocovanie jednotlivých typov zberov (tematický, výberový, komplexný). Kurátori vytvárajú zoznamy a nastavujú konfiguračné parametre Heritrixu pre zber, vyberajú profily, na základe ktorých sa definuje hĺbka zberu. V prípade potreby vedia konkrétne stránky zablokovať. Regulujú prácu workerov, spúšťajú indexáciu, vedia odhaliť pasce, ktoré spôsobujú zacyklenie zberu. Kurátori dopĺňajú a editujú metadáta v katalógu webových stránok i v katalógu e-Born prameňov a rozhodujú o tom, čo a v akom rozsahu (v súlade s uzatvorenými zmluvami) sprístupnia koncovým používateľom.



Obr. 4 Administrátorské rozhranie pre kurátorov

Rutinná prevádzka – rok 1

Rutinná prevádzka IS Digitálne pramene bola spustená 1. januára 2016. Oddelenie DDP pripravilo výberový zber pri príležitosti konania parlamentných volieb do Národnej rady SR (samotné voľby sa konali dňa 5. marca 2016). Zber mapoval predvolebnú kampaň, volebné výsledky, reakcie na ne i povolebný vývoj až do zostavenia novej vlády. Zbieralo sa 220 webových sídiel politických strán, kandidátov a lídrov politických strán, webové stránky prieskumných agentúr a vybrané blogy. Zber prebiehal od januára do konca marca v týždenných intervaloch, po voľbách sa frekvencia zberov upravila na dvakrát týždenne. Z celkového počtu 220 sa podarilo úspešne zozbierať 190 zdrojov.

V súčasnosti sa pripravuje tematický zber webových stránok kultúrnych inštitúcií, ktorý bude vychádzať z pilotného tematického zberu, avšak v užšom zábere. Prioritne sa bude orientovať na inštitúcie, ktoré spadajú do pôsobnosti Ministerstva kultúry SR (knižnice, galérie, múzeá).

V rámci vyššie spomínaného pilotného zberu začali kurátori Depozitu digitálnych prameňov oslovovať vybrané inštitúcie z tematickej oblasti Kultúrny profil Slovenska. Od 6. novembra 2015 do 30. apríla 2016 sa oslovilo vyše 400 inštitúcií a uzatvorilo 80 zmlúv na 107 URL adries. Z tohto počtu bolo 95 webových stránok a 12 e-Born titulov, všetko elektronické seriály.

V druhom polroku 2016 sa plánuje realizovať výberový zber k Predsedníctvu SR v Európskej únii. Príprava na zber zahŕňa výber semienok (URL adries) a ich testovanie.

Aktuálne sa veľká pozornosť venuje problematike e-Born dokumentov. V prvej etape sa riešia elektronické seriály, ku ktorým neskôr pribudnú aj elektronické monografie. Pozornosť sa venuje predovšetkým typológii e-Born seriálov, nakoľko každý titul vyžaduje individuálny prístup. V súlade s Politikou zberu e-Born seriálov sa týmto termínom označuje „individuálny informačný prameň na pokračovanie, ktorý je jednoznačne identifikovaný názvom, identifikátorom a metadátami" [6]. Analýza jednotlivých typov elektronických seriálov je dôležitá pre správne nastavenie postupov prác s týmto typom dokumentov (spracovanie, ukladanie do archívu, tvorba metadátových záznamov v MARC 21 a sprístupňovanie). Všetky typy elektronických seriálov musia spĺňať kritériá pre pridelenie ISSN čísla. Výber prameňov pre zber a archiváciu preto prebieha v úzkej spolupráci s oddelením NA ISSN. Obdobne pri výbere monografií sa budú vyberať predovšetkým on-line monografie s prideleným ISBN číslom, pričom sa bude spolupracovať s NA ISBN.

Legislatívna situácia

Aktuálna legislatívna situácia na Slovensku poskytuje iba obmedzené možnosti pre archiváciu a sprístupnenie archivovaného obsahu z archívu. V súlade s platným Autorským zákonom [7] je možné vyhotoviť rozmnoženinu pre archívne účely iba z diela z vlastných fondov. Jej sprístupňovanie v zákone upravené nie je. Najjednoduchšia situácia je pri prameňoch označených verejnými licenciami typu Creative Commons a pri Open Access tituloch. V prípade elektronických prameňov označených licenciami Creative Commons nie je potrebné uzatvárať zmluvy s vydavateľmi a ukladanie do archívu i sprístupňovanie sa robí na základe tejto licencie.

Z hľadiska ochrany autorských práv je v ostatných prípadoch potrebné individuálne uzatvárať zmluvy s vydavateľmi, resp. poskytovateľmi. Zmluva o poskytovaní elektronických on-line prameňov bola pripravená v rámci intenzívnej spolupráce s právnikom, odborníkom na autorské právo. Je vystavená na stránke www.webdepozit.sk, odkiaľ si ju záujemcovia môžu stiahnuť a doplniť potrebné údaje. Uzatvorenie zmluvy prebieha výlučne na báze dobrovoľnosti zo strany poskytovateľa (vydavateľa), ktorý sám rozhoduje o tom, čo umožní zbierať (čiže archivovať) resp. sprístupniť a v akom režime. V súčasnosti sú možné tri typy prístupu k prameňom v archíve: voľný prístup (neobmedzený, odkiaľkoľvek), lokálny (obmedzený na vybrané počítače v priestoroch Univerzitnej knižnice v Bratislave) a bez prístupu (prístup zakázaný).

On-line publikácie nie sú v aktuálne platnom Zákone o povinnom výtlačku [8] vôbec zadefinované. Povinnosť vydavateľov odovzdávať povinný výtlačok sa vzťahuje iba na elektronické publikácie na fyzických nosičoch (offline). V blízkej budúcnosti bude preto potrebná legislatívna úprava inštitútu povinného výtlačku a rozšírenie jeho záberu o elektronické on-line pramene tak, ako je to dnes už bežnou praxou v mnohých krajinách Európy i sveta.

Význam našich aktivít spočíva najmä v získavaní skúseností so spracovaním rôznych typov elektronických prameňov, v otestovaní jednotlivých postupov a procesov a vo vytvorení metodiky práce s týmito typmi dokumentov. Tým sa vytvoria praktické predpoklady pre zvládnutie elektronického povinného výtlačku v našich podmienkach.

Záver

Na základe uvedených skutočností a skúseností z realizácie projektu, ako aj z komplexnosti riešenia sa dá konštatovať, že ide o vysoko náročný a na Slovensku jedinečný projekt. Prevádzka softvérových systémov, hardvéru i organizačno-metodické zabezpečenie rutinnej prevádzky nie sú triviálne a budú si vyžadovať kontinuálnu údržbu, vývoj a aktualizáciu metodických postupov. Veľa bude záležať aj na doriešení legislatívnych otázok, dlhodobého zabezpečenia finančných prostriedkov a ľudských zdrojov. Ambíciou Univerzitnej knižnice v Bratislave aj naďalej zostáva cieľ dlhodobo udržať tento systém v prevádzke a ochrániť tak digitálne kultúrne dedičstvo pre nasledujúce generácie.

Použitá literatúra

[1] *Správa o činnosti a hospodárení za rok 2015* [online]. Bratislava: Univerzitná knižnica v Bratislave, 2015 [cit. 2016-05-03]. Dostupné na: http://www.ulib.sk/files/vyr-sprava/sprava_cinnosti_ukb_2015.pdf

[2] KONEČNÝ, Michal, 2016. *Návrh kompetenčního modelu a kurikula digitálního kurátorství*. *Diplomová práce* [online]. Brno: Masarykova univerzita, Filozofická fakulta, 2016 [cit. 2016-05-05]. Dostupné na: http://is.muni.cz/th/426710/ff_m/Michal_Konecny_-_Kompetencni_model_a_kurikulum_digitalniho_kuratorstvi_-_Diplomova_prace_-_prilohy.pdf

[3] *What is digital* curation? [cit. 2016-05-05]. Dostupné na: http://www.dcc.ac.uk/digital-curation/what-digital-curation

[4] *Preservation Management of Digital Materials: Handbook* [online]. Digital Preservation Coalition, 2008 [cit. 2016-05-05]. Dostupné na: http://www.dpconline.org/component/docman/doc_download/299-digital-preservation-handbook

[5] MATÚŠKOVÁ, Jana, 2015. *Politika zberu DIP www* [online]. [cit. 2016-05-09]. Dostupné na: https://www.webdepozit.sk/sk/projekt-dip/dokumentacia/zoznam-dokumentov

[6] KATRINCOVÁ, Beáta, 2015. *Politika zberu DIP e-Born seriály* [online]. [cit. 2016-05-09]. Dostupné na: https://www.webdepozit.sk/projekt-dip/dokumentacia/zoznam-dokumentov

[7] *Zákon č. 185/2015 Z. z. – Autorský zákon*

[8] *Zákon č. 212/1997 Z. z. o povinných výtlačkoch periodických publikácií, neperiodických publikácií a rozmnožením audiovizuálnych diel v znení neskorších predpisov.*

Ing. Alojz Androvič, PhD. Alojz.androvic@ulib.sk
Mgr. Ivan Ciglan ivan.ciglan@ulib.sk
PhDr. Jana Matúšková *jana.matuskova@ulib.sk*
(Univerzitná knižnica v Bratislave)