Uchovávanie digitálnych objektov a ich súčastí v prostrediach open-source systémov Invenio a DSpace

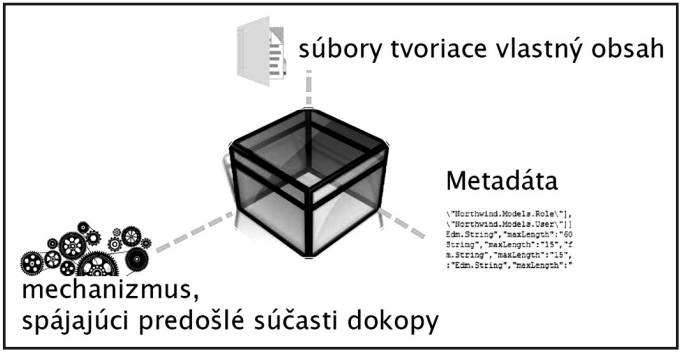
V predkladanom odbornom príspevku rozoberáme vybrané aspekty vonkajšej a vnútornej reprezentácie súčastí digitálnych objektov vo vybraných dvoch open-source systémoch. Do týchto systémov sme počas experimentu importovali zvolenú identickú elektronickú kolekciu, pozostávajúcu z viacerých textových, ako aj multimediálnych súborov. Následne v príspevku vzájomne, medzi skúmanými systémami, porovnávame metódy postupného prideľovania statických, vonkajších identifikátorov k novovzniknutým digitálnym objektom, v kontraste s dátovými štruktúrami ich systémovej vnútornej reprezentácie priamo v súborovom systéme unix-like servera.

Digitálny objekt

Podľa Bartoška (2004), môžeme v širšej významovej rovine definovať digitálny objekt ako dátovú štruktúru pre základnú, samostatne použiteľnú informačnú jednotku. Z pohľadu elektronického systému zas digitálny objekt vnímame tiež ako počítačom spracované a spracovávané údaje umiestnené spoločne s príslušnými metadátami v určitom digitálnom priestore a kontexte.

V odbornej akademickej príručke *Guidelines for Digital Objects* (University of California, 2011) nachádzame definíciu, podľa ktorej, digitálne objekty pozostávajú z troch primárnych komponentov:

- jedného alebo viacerých elektronických súborov tvoriacich vlastný obsah,
- príslušných metadát,
- mechanizmu, ktorý asociuje predošlé dve spomínané súčasti dokopy.



Obrázok 1 Vizualizácia rozmerov digitálneho objektu podľa definície z University of California (2011)

Vychádzajúc z vyššie uvedenej definície sme na Obrázku 1 ilustrovali jednotlivé súčasti digitálneho objektu s využitím troch pomyselných dimenzií.

Neodmysliteľnou súčasťou digitálneho objektu sú jeho metadáta. Podľa Bartoška (2004) vnímame metadáta (metaúdaje) v kontexte digitálnych knižníc „ako počítačovo spracovateľné štruktúrované informačné objekty popisujúce vlastnosti iných informačných objektov." Spojenie metadát so súbormi digitálneho obsahu v jeden funkčný celok považujeme v kontexte digitálnych knižníc a repozitárov za kľúčový moment.

Saidis a Delis (2007) vo svojom odbornom článku definujú digitálny objekt v kontexte nastolenej problematiky cez jeho štyri charakterizujúce elementy: obsah objektu, jeho metadáta, vzťahy s inými objektmi, a napokon cez správanie. Niektorými vybranými aspektmi digitálnych objektov sa budeme zaoberať aj v tomto príspevku.

Motivácia k uskutočneniu tohto druhu experimentu vyplýva z existencie viacerých typov open-source systémov na podporu digitálnych knižníc a repozitárov, pričom vnímame medzi týmito systémami výskyt značných rozdielov, okrem iných aspektov, aj v oblasti metód identifikácie a dlhodobého uchovávania samotných digitálnych objektov.

Interný pohľad na súborový systém digitálneho repozitára sa ponúka zvyčajne len úzkej množine používateľov-odborníkov či systémových správcov. Možno aj to je príčinou nedostatočného preskúmania spomínanej problematiky a absencie odpovedajúcich aktuálnych komparatívnych analýz. Kvalitne vypracované analýzy k tejto téme pochádzajú spred niekoľkých rokov a bolo by potrebné ich aktualizovať. Za všetky spomeňme aspoň odbornú analýzu *Digital libraries: Comparison of 10 software* (Andro et al., 2012), v ktorej nájdeme, okrem iného, aj komplexné porovnanie starších verzií systémov Invenio 1.0.0 a DSpace 1.7.2.

Sme presvedčení, že administrátori všetkých elektronických systémov potrebujú detailne poznať ich správanie, aby dokázali pružne reagovať na rôznorodé situácie vznikajúce pri ich správe. Spresniť pohľad na digitálny objekt a jeho súčasti v dvoch vybraných systémoch môžu aj výsledky experimentu, ktoré prinášame v tomto príspevku.

Cieľ a metodika experimentu

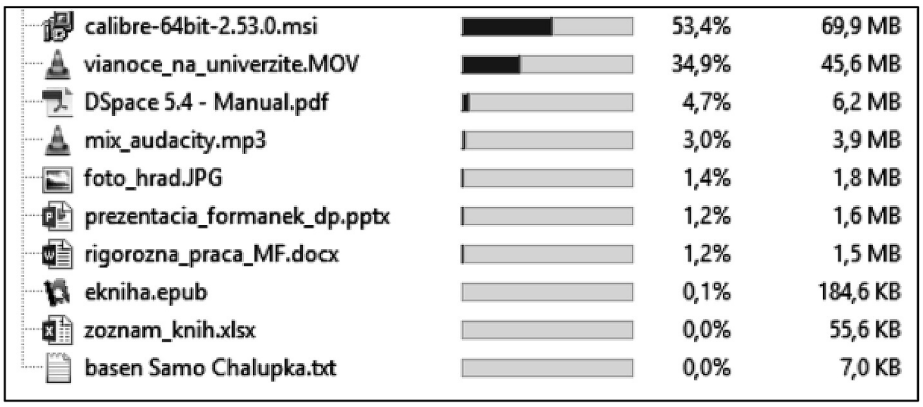
Cieľom experimentu bola vzájomná komparácia spôsobov, ktorými dochádza ku vnútornej reprezentácií digitálnych objektov a ich označeniu pod aplikačnou vrstvou vybraných open-source systémov na podporu digitálnych knižníc a repozitárov. Sledovali sme taktiež spôsoby identifikácie digitálnych objektov z vonkajšieho pohľadu, tak ich vníma aj používateľ konkrétneho systému.

K uskutočneniu experimentu sme mali k dispozícii dva samostatné nezávislé open-source systémy: DSpace 5.5 a Invenio 1.1.6. Oba systémy boli nainštalované na samostatných serveroch s identickými výkonnostnými parametrami v rámci virtuálnej infraštruktúry. Ako základný serverový operačný systém sme zvolili Ubuntu server 14.04.4 LTS, teda verziu s dlhoročnou podporou.

Aby sme zabezpečili riadené a identické podmienky priebehu celého experimentu, do oboch systémov bolo nutné vložiť identické digitálne objekty, resp. cielene zvolenú množinu digitálnych súborov s reálnym obsahom, spoločne s príslušnými metadátami v zmysle už spomínanej definície[1].

Starostlivo sme teda volili testovaciu množinu elektronických súborov, ktoré sme následne postupne importovali do každého z oboch skúmaných systémov. Pri výbere vzorky testovacích súborov sme sledovali tieto podmienky:

- Počet digitálnych súborov musí byť identický pre oba systémy. Kvôli efektívnejšej manipulácii sme obmedzili ich počet na desať.
- Aby bola vzorka variabilnejšia, mala by pokrývať širší záber tých najbežnejších a najpoužívanejších formátov, vrátane multimédií.
- Súbory by mali byť voľne dostupné.

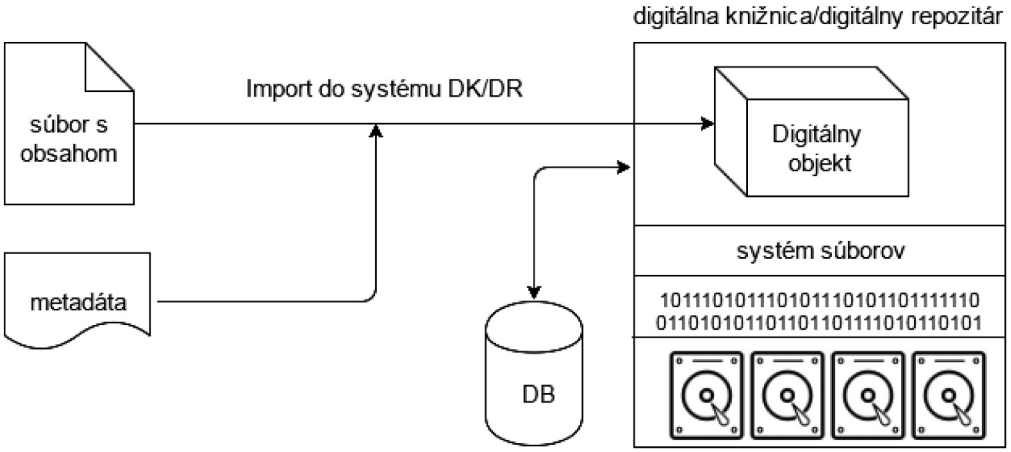


Obrázok 2 Množina testovacích digitálnych súborov

Berúc do úvahy predošlé kritériá, zvolili sme množinu desiatich testovacích súborov tak, ako ju vidíme na Obrázku 2. V tejto množine sú zastúpené dokumenty textového charakteru (súbor formátu PDF, dokument Word, zošit programu Excel, elektronická kniha vo formáte EPUB, prostý formát obyčajného textu TXT), ako aj bežné multimediálne súbory. Všetky dokumenty pritom obsahujú reálne informácie odpovedajúceho charakteru. Do vzorky sme začlenili aj spustiteľný inštalátor open-source softvéru na správu osobnej zbierky elektronických kníh. Súhrnný objem digitálnych materiálov presiahol 130 MB.

[1] Viď. Guidelines for Digital Objects (University of California, 2011)

K týmto zvoleným súborom digitálneho obsahu sme v momente importu do systémov Invenio a DSpace priradili i najzákladnejšie metadáta: názov, meno autora a jedno kľúčové slovo. Od tohto momentu ich teda v zmysle definície[2] môžeme vnímať pod všeobecnejším pojmom *digitálny objekt*. Proces vzniku digitálneho objektu sme sa pokúsili zachytiť na Obrázku 3.



Obrázok 3 Vznik digitálneho objektu

Digitálne objekty v Invenio 1.1.6

Úvodná inštalácia systému umožňovala importovať tzv. *demo records*, teda ukážkové digitálne objekty, ku ktorým sme, v zmysle metodiky experimentu, pridali aj vybrané testovacie objekty spoločne s metadátami. V tomto prípade sme využili dávkový import[3] prostredníctvom webového rozhrania systému priamo do novovytvorenej kolekcie *TEST*.

V ďalšom kroku sme sa pomocou terminálu pripojili k serveru Invenio a pristúpili ku skúmaniu fyzických súborov zapísaných na pevnom disku unix-like servera. V zmysle aktuálne platnej konfigurácie, boli všetky súbory umiestnené v prednastavenom priečinku /opt/invenio/var/data/files/g0/, v ktorom začínala samotná adresárová štruktúra uložených digitálnych súborov. Zistili sme, že táto štruktúra obsahovala adresáre uložené v jednej úrovni[4], pričom boli označené vzostupne číselným označením „1" až „244". Značná časť adresárov obsahovala *demo-records*, teda jednotlivé súčasti množiny digitálnych objektov, ktoré boli v skúmanom systéme prítomné, hneď po inštalácii. K týmto objektom sa sekvenčne priradila aj naša vzorka objektov.

Pri širšom prehľadávaní štruktúry adresárov absolútnej cesty /opt/invenio/var/data/files/g0**/** sme zistili, že *demo-records,* siahajú od adresára „1" po adresár s číslom „234". Nami importované testovacie súbory teda našli miesto až na konci adresárovej štruktúry – v desiatich priečinkoch označených vzostupne číslami „235" až „244". Opätovne sme skúmali túto štruktúru s cieľom identifikovať obsah adresárov.

```
   6517 mar 26 09:19 Dspace 5_4 Manual.gif;icon;1
   8864 mar 26 09:19 Dspace 5_4 Manual.jpg;icon-180;1
  32780 mar 26 09:19 Dspace 5_4 Manual.jpg;icon-700;1
6460278 mar 26 09:14 Dspace 5_4 Manual.pdf;1
    255 mar 26 09:19 .md5
      3 mar 26 09:14 .recid
1611055 mar 26 09:20 .text;1
      4 mar 26 09:14 .type
```

Obrázok 4 Súbory tvoriace časť digitálneho objektu

Ukázalo sa, že každý číselne označený adresár obsahuje časti samostatného digitálneho objektu – predovšetkým súbor s vlastným obsahom, pričom ten je fyzicky uložený v špeciálnej forme a podľa špecifických pravidiel, ako je zrejmé aj z Obrázku 4. Vidíme tu zobrazený obsah zložky, ktorej je uložený jeden z našich testovacích súborov – manuál vo formáte PDF.

Názov súboru je odlišný od toho, pod ktorým sme súbor importovali. Pre interné účely systému je súbor s vlastným obsahom pomenovaný ako *DSpace 5_4 Manual.pdf;1*. Okrem neho sú v adresári prítomné aj ďalšie pomocné súbory.

Zistili sme, že v skrytom súbore *.recid*[5] je uložená informácia o pridelenom rovnomennom identifikátore nazvanom *recID*, ktorý identifikuje konkrétny záznam a automaticky sa prideľuje záznamom inkrementálne (CERN, 2015).

Obsah textového súboru *.type* prezrádza, že ide o hlavný súbor záznamu, čo dokazuje prosté slovo *Main* uložené vo vnútri tohto textového súboru.

---

[2] Uvedená v Guidelines for digital objects (The University of California, 2011)

[3] Problematika importu v systéme Invenio je rozsiahla a presahuje priestor tohto príspevku. Navyše, v jeho kontexte nie je našim cieľom momentálne podrobnejšie skúmať dávkový import digitálnych objektov.

[4] Všetky adresáre sa nachádzali na rovnakej úrovni stromu, pričom vo vnútri adresárov sa už ďalšie priečinky nevyskytovali.

[5] Úvodná bodka je pevnou súčasťou názvu skrytého súboru. Ide o bežnú konvenciu zaužívanú v prostrediach unix-like operačných systémov.

Na Obrázku 4 vidíme ďalej, že systém vytvoril aj viacero typov ikon, spočítal MD5 hash vkladaného súboru a vyextrahoval text z PDF kvôli plnotextovému prehľadávaniu tohto digitálneho objektu. Takto získaný text následne systém Invenio uložil do ďalšieho skrytého súboru *.text;1*.

V ďalšej časti experimentu sme cez webové rozhranie Invenio pridali do už existujúceho záznamu s recID 257 ďalší, sekundárny PDF súbor, označený pri importe za doplnkový (additional). Túto experimentálnu časť sme vykonali za účelom zistenia, akým spôsobom Invenio doplnkový súbor spracuje a uloží sa do existujúcej štruktúry.

Preskúmali sme obsah jedenástich naposledy pridaných priečinkov (10 základných + 1 doplnkový súbor), nachádzajúcich sa v adresári /opt/invenio/var/data/files/g0**/.** Pre lepšiu prehľadnosť sme zistené skutočnosti zapísali do Tabuľky 1.

P.č.	Názov fyzického súboru uloženého v adresári na .../g0/	Pridelené recID	Číslo adresára v .../g0/	Extrahovaný text z pôvodného súboru	Vytvorené ikony	Obsah súboru *.type*
1.	Calibre 253.msi;1	255	235	Nie	Nie	Main
2.	Báseň Samo Chalupka.txt;1	254	236	Áno	Nie	Main
3.	DSpace 5_4 Manuál.pdf;1	257	237	Áno	Gif; jpg 180;jpg 700	Main
4.	Foto hradu.jpg;1	259	238	Nie	Gif; jpg 180;jpg 700	Main
5.	Mix vytvorený v Audacity.mp3;1	260	239	Nie	Nie	Main
6.	Prezentácia k diplomovej práci.pptx;1	261	240	Nie	Nie	Main
7.	Využitie digitálnej knižnice v akademickom prostredí Žilinskej univerzity.docx;1	262	241	Nie[6]	Nie	Main
8.	Vianoce na univerzite.MOV;1	263	242	Nie	Nie	Main
9.	Zoznam kníh na katedre.xlsx;1	264	243	Nie	Nie	Main
10.	Ekniha.epub;1	258	244	Nie	Nie	Main
+ doplnkový PDF súbor						
Manuál Invenio.pdf;1		257	245	Áno	Gif; jpg 180;jpg 700	Additional

Tabuľka 1

[6] Domnievame sa, že ide len o nedokonalosť v aktuálnej konfigurácii systému. Extrakciu textu z formátu DOCX je nutné povoliť dodatočne.

Z údajov uvedených v Tabuľke 1 a zistených skutočností vyplýva, že:

- systém Invenio radí všetky novopridané digitálne objekty postupne na koniec adresárovej štruktúry v priečinku[7] určenom na fyzické ukladanie súborov.
- Pôvodné pomenovania súborov sú nahradené názvom definovaným pri procese importu, ku ktorému je pridaný sufix „ ;1". Ako sme si však overili, tento sufix nevyjadruje poradie súborov vlastného obsahu v konkrétnom digitálnom objekte.
- Ukázalo sa, že ak pridáme k už existujúcemu digitálnemu objektu (do záznamu s tým istým recID) ďalší súbor vlastného obsahu, tento bude na pevnom disku uložený úplne samostatne a nezávisle (viď koniec Tabuľky 1), hoci v aplikačnej úrovni budú tvoriť oba súbory pevnú súčasť toho istého digitálneho objektu.
- Prideľovaný identifikátor recID je nezávislý na pomenovaní adresára s vlastným obsahom. Táto skutočnosť môže byť pre používateľa irelevantná, nakoľko sa prostredníctvom webového rozhrania prezentuje digitálny objekt len pod číslom recID.
- Systém Invenio zo špecifických formátov súborov tvoriacich obsah (minimálne TXT, PDF – po zmene konfigurácie aj ďalšie) automaticky na pozadí extrahuje text a umiestňuje ho do separátneho textového súboru, čím umožňuje používateľom plnotextové vyhľadávanie naprieč digitálnymi objektmi.
- Invenio automaticky vytvára ikony pre PDF a JPEG súbory, a to paralelne v dvoch formátoch JPEG a GIF, navyše v dvoch rôznych rozlíšeniach obrazu.
- Invenio dbá na kontrolu integrity digitálnych objektov pomocou funkcie MD5. Výsledok je vypočítaný pre každý vzniknutý čiastkový súbor zvlášť (vrátane ikon) a uložený v osobitnom súbore *.md5* v príslušnej zložke digitálneho objektu.
- Uvedené súčasti uložených digitálnych objektov prepájajú a dopĺňajú údaje zapísané v tabuľkách použitej relačnej databázy typu MySQL.
- Úložisko (v našom prípade adresár /opt/invenio/var/data/files/g0/) je spoločne s obsahom príslušnej databázy systému kľúčovým prvkom pri zálohovaní.
- Obsah úložiska je čitateľný aj bez informácií z databázy, napríklad v krízovom prípade jej nedostupnosti.

Digitálne objekty v DSpace 5.5

Druhým testovaným systémom, ktorému sa budeme v tomto príspevku venovať, je open-source softvér na podporu digitálnych repozitárov s názvom DSpace. Spustená inštancia systému neobsahovala, na rozdiel od Invenio, žiadne vývojármi dodané záznamy. Z tohto dôvodu bolo nutné pre uskutočnenie testu inicializovať systém a definovať v ňom minimálne jednu komunitu s integrovanou kolekciou pre budúce digitálne objekty. Pre účely experimentu sme vytvorili testovaciu katedrovú komunitu *KMKD* a kolekciu *TEST.*

Vývojári systému DSpace uvádzajú, že jednou z kľúčových vlastností tohto systému je priradenie trvalého, perzistentného identifikátora pre každú položku obsahu digitálneho repozitára, vrátane kolekcií digitálnych objektov a im nadradených komunít. Jedným z perzistentných identifikátorov je tzv. *handle*. Inštancia DSpace môže, po príslušnej registrácii, k priradzovaniu a udržiavaniu *handles* využívať systém s názvom *CNRI*[8] *Handle System* (The DSpace Developer Team, 2016, s. 26, s. 600).

Problematika registrovania *handle* prefixov z CNRI a následné používanie identifikátorov presahuje možnosti tejto práce a nie možné sa momentálne tejto téme venovať do väčšej hĺbky. Pre naše účely momentálne postačí skutočnosť, že digitálnym objektom, kolekciám i komunitám, sú pomocou špecializovaného systému priradzované trvalé *handle* identifikátory.

Z webového používateľského rozhrania DSpace sme zistili, že komunite *KMKD* bol pridelený identifikátor hdl:123456789/1. Testovacej kolekcii s názvom *TEST*, systém automaticky pridelil hdl:123456789/2. Vidíme, že prefix[9] použitý pred lomkou ostáva rovnaký pre všetky entity použité v systéme. Na základe dostupných informácií získaných z manuálových stránok usudzujeme, že pomocou *handle* budú označené aj ďalšie digitálne objekty, ktorých súbory obsahu sme importovali v ďalšom kroku porovnávacieho experimentu.

Od „vonkajšieho", perzistentného označenia digitálnych objektov pomocou *handle* identifikátora, presuňme našu pozornosť k vnútornej reprezentácii objektov uchovávaných v systéme DSpace. V tejto časti príspevku zhrnieme spôsoby ukladania fyzických súborov digitálnych objektov. Systém DSpace využíva pre tento účel osobitný adresár, tzv. *asset store*. Ten pozostáva z jedného alebo viacerých koreňových priečinkov so známym a prispôsobiteľným umiestnením. V priečinku, resp. priečinkoch tvoriacich *asset store* vzniká priestor pre ďalšiu štruktúru, ktorú sme chceli bližšie preskúmať.

V zmysle aktuálnej konfigurácie bol v skúmanom repozitári DSpace vytvorený len jeden *asset store* a nachádzal sa v rovnomennom priečinku na adrese /dspace/assetstore/. Manuál k DSpace (The DSpace Developer Team 2016) uvádza, že nízkoúrovňovú správu týchto úložísk zabezpečuje softvérová vrstva označovaná ako *Bitstream Storage Manager* (ďalej len BSM), ktorá koordinuje ukladanie jednotlivých súčastí digitálnych objektov v podobe tzv. *bitstreams.* Ide o súbory uložené v *asset store,* ktorých pôvodný názov bol pozmenený na 38-miestny interný číselný identifikátor (The DSpace Developer Team, 2016, s. 745). Toto označenie nie je viditeľné zvonku, pretože ho používa výhradne BSM na presné uloženie a vyhľadanie súboru v adresárovej štruktúre. Predmetnú štruktúru a jej vznik v konkrétnom *asset store* sme skúmali pri postupnom importe desiatich digitálnych objektov.

---

[7] v našom prípade /opt/invenio/var/data/files/g0/. Absolútna cesta umiestnenia sa môže líšiť v závislosti od konfigurácie systému.

[8] Skratka z anglického Corporation for National Research Initiatives. Viac informácií získa čitateľ na http://www.handle.net/

[9] Uvedené číselné označenie prefixu nie je registrované v CNRI, preto systém používa prednastavenú, testovaciu hodnotu 123456789.

Zistili sme, že pred importom súborov bola adresárová štruktúra úplne prázdna. Rovnako, ako pri systéme Invenio, sme vložili i do inštancie DSpace 10 základných plus 1 doplnkový testovací súbor (viď Obrázok 2). Každý z nich bol popísaný len minimálnym súborom metadát.

Následne sme sledovali, akým spôsobom systém vzniknuté digitálne objekty globálne identifikuje a ako uloží potrebné súbory do štruktúry v *asset store*. Zistené výsledky prinášame zosumarizované v Tabuľke 2. Stĺpec tejto tabuľky pomenovaný „Uložené bitové toky"[10], označuje fyzické súbory obsahu pripojené ku konkrétnemu záznamu. *Bitové toky* korešpondujú s už spomínanými *bitstreams*, avšak interne používaný číselný identifikátor je, pomocou údajov z PostgreSQL databázy, nahradený pôvodným názvom súboru, ktorým bol súbor pomenovaný v čase jeho importu do systému. Z tohto dôvodu budeme v kontexte tohto príspevku za *bitové toky* považovať importované elektronické súbory s menným označením (vrátane prípony vyjadrujúcej formát), začlenené do príslušného záznamu reprezentovaného identifikátorom *handle*.

P.č.	**Názov objektu uložený v metadátach**[11]	**Pridelený *handle* identifikátor**	**Uložené *bitové toky***	**Automaticky rozpoznaný formát *bitového toku***
1.	Mix Audacity	123456789/3	Mix_audacity.mp3	neznámy
2.	Báseň Samo Chalupka	123456789/4	Basen Samo Chalupka.txt	Textový súbor
3.	Calibre 2.53	123456789/5	Calibre-64bit-2.53.msi	neznámy
4.	Ekniha	123456789/6	Ekniha.pub	Neznámy
5.	DSpace 5.4 Manuál	123456789/7	DSpace 5.4 – Manual.pdf	Adobe PDF
6.	Foto hrad	123456789/8	Foto_hrad.jpg	JPEG
7.	Prezentácia k diplomovej práci	123456789/9	prezentacia_formanek_dp.pptx	Microsoft PowerPoint
8.	Využitie digitálnej knižnice v akademickom prostredí Žilinskej univerzity	123456789/10	Rigorozna_praca_MF.docx	Microsoft Word
9.	Vianoce na Žilinskej univerzite	123456789/11	Vianoce_na_univerzite.MOV	Video Quicktime
10.	zoznam kníh v knižnici KMKD	123456789/12	Zoznam_knih.xlsx	Microsoft Excel
Doplnkový súbor:				
Manuál Invenio		123456789/7	Manual_Invenio.pdf	Adobe PDF

Tabuľka 2 Vonkajšia identifikácia digitálnych objektov v DSpace

[10] Označenie „bitové toky" je prevzaté z oficiálne dostupnej českej lokalizácie webového prostredia XMLUI pre systém DSpace 5.4

[11] Ide o názov uložený v zázname Qualified Dublin Core pod dc.contributor.author

Ku záznamu hdl:123456789/7 sme pridali doplnkový súbor PDF, ktorý sa stal pevnou súčasťou tohto digitálneho objektu. „Vonkajší", používateľský pohľad na takto vzniknutý objekt s dvomi pridruženými bitovými tokmi môžeme vidieť na Obrázku 5. K obom tokom je v detailoch záznamu vypočítaná hodnota funkcie MD5. Systém DSpace, podobne ako Invenio, dbá týmto spôsobom na integritu uložených digitálnych objektov.

Obrázok 5 Dva súbory PDF ako súčasť jediného digitálneho objektu s hdl: 123456789/7

Pri analýze predmetného *asset store* sa ukázalo, že desať vložených súborov bolo uložených na disk servera v podobe desiatich *bitstreams* rozmiestnenými rovnomerne pod tromi úrovňami dvojciferne pomenovaných adresárov. Túto skutočnosť sme zachytili v Obrázku 6. Z Obrázka 6 je teda zrejmé, že skutočne došlo k prepísaniu pôvodného názvu každého importovaného súboru 38 ciframi. Prvých šesť znakov číselného označenia pritom zodpovedá umiestneniu *bitstream*-u v trojúrovňovej adresárovej štruktúre. Prvých šesť znakov tak určuje tri vetvy stromovej štruktúry, v ktorej je následne každý *bitstream* uložený, ako list vznikajúceho stromu.

Na Obrázku 6 sme modrou farbou zvýraznili dodatočne identifikovaný doplnkový súbor PDF (Manual_Invenio.pdf), importovaný ako jedenásty v celkovom poradí (viď koniec Tabuľky 2). Vidíme, že BSM preň vytvoril osobitné umiestnenie v relatívnej ceste 86/03/30/. Zmenu v adresárovej štruktúre *asset store*, ktorá vznikla pridaním doplnkového súboru, sme zachytili metódou porovnania dvoch kópií[12] *asset store* – pred importom a po importe doplnkového súboru. Skutočnosť, že ide naozaj o importovaný PDF manuál, sme overili skopírovaním *bitstream* súboru z lokálneho disku servera. Nedbajúc na číselný názov, sme v systéme Windows doplnili príponu .pdf. Súbor sa následne podarilo bez problémov otvoriť asociovaným softvérom (Napr. Foxit Reader). Vedeli sme, že ako posledný súbor sme importovali dokument formátu PDF. Ak by sme však nedisponovali touto informáciou, formát súboru *bitstream* jednoduchým spôsobom nezistíme, pretože číselné označenie nevypovedá nič o pôvodnom formáte a názve súboru. Z manuálu k DSpace 5.x (The DSpace Developer Team, 2016, s. 742) je zrejmé, že tieto údaje sú uložené výhradne v tabuľkách relačnej databázy spoločne s inými metadátami.

Z vyššie uvedených poznatkov a experimentálne overených skutočností vyplýva, že:

Názov súboru	Adresár	Výsledok porovnania	Veľkosť vp...	Veľkosť vľavo
86		Iba vpravo: C:\Users\matus_0...		
03	86	Iba vpravo: C:\Users\matus_0...		
30	86\03	Iba vpravo: C:\Users\matus_0...		
860330073532005587019587481302085187 56	86\03\30	Iba vpravo: C:\Users\matus_0...	2,52 MB	
52039949107918395938552430783303993804	52\03\99	Binárne súbory sú identické	1,53 MB	1,53 MB
58509537983551749927661083475884177069	58\50\95	Binárne súbory sú identické	55,6 KB	55,6 KB
60259545901879551283418885113682198151	60\25\95	Binárne súbory sú identické	1,61 MB	1,61 MB
72024124885483201941883854405742925 05	72\02\41	Binárne súbory sú identické	45,6 MB	45,6 MB
12500707107394858216412283000810375965 6	12\50\07	Binárne súbory sú identické	1,83 MB	1,83 MB
14614802700600551656428651035909738345 2	14\61\48	Binárne súbory sú identické	69,9 MB	69,9 MB
47390512613209764394557999915580633218	47\39\05	Binárne súbory sú identické	6,16 MB	6,16 MB
50896038484317754630705826479595144286	50\89\60	Binárne súbory sú identické	185 KB	185 KB
99647227373340241531937893764092154075	99\64\72	Binárne súbory sú identické	3,87 MB	3,87 MB
32246436272273401288748437910929375750	32\24\64	Text files are identical	7,05 KB	7,05 KB

Obrázok 6 Vnútorné reprezentácie bitových tokov v systéme DSpace

- Každému novému záznamu, kolekcii aj komunite je systémom DSpace vzostupne priradený identifikátor *handle*, ktorý slúži na identifikáciu entity pre potreby používateľov, či iných kooperujúcich systémov. *Handle prefix* je potrebné zaregistrovať, inak sa vždy použije testovacia hodnota.
- K jednej entite označenej pomocou *handle* identifikátora môže prislúchať viacero *bitových tokov* (a teda aj *bitstreams*), ktoré sú ale v súborovom systéme uložené úplne nezávisle (rovnako ako v prípade softvéru Invenio).
- DSpace natívne správne rozpoznáva bežné formáty importovaných elektronických súborov: DOCX, XLSX, PDF, TXT, JPEG, MOV.

[12] Využili sme softvér WinMerge, ktorý je primárne určený na hľadanie rozdielov medzi dvomi adresármi.

- Subsystém BSM, ktorý je súčasťou DSpace využíva vlastné štruktúry pre interné účely manipulácie so súčasťami digitálnych objektov: vkladané elektronické súbory premenováva 38-ciferným číselným identifikátorom. Na základe prvých šiestich znakov umiestňuje BSM tieto *bitstreams* do trojvrstvovej adresárovej štruktúry v príslušnom *asset store*. Podľa vyjadrení vývojárov je zmyslom tejto konvencie zvýšenie efektivity pri fyzickom pristupovaní k týmto súborom (The DSpace Developer Team, 2016, s. 745)
- Číselné pomenovanie pre každý nový *bitstream* sa generuje zrejme s využitím hashovacej funkcie, pričom vygenerované pomenovanie musí byť unikátne v rámci systému.
- Výhodou náhodného generovania číselných pomenovaní pre *bitstreams* je rovnomerné rozloženie súborov naprieč celým úložiskom *asset store.*
- Integritu všetkých bitových tokov zabezpečuje funkcia MD5. O pravidelnú kontrolu integrity obsahu *asset store* dbá používateľ plánovaním vykonávania príslušného skriptu naprieč úložiskom.
- Tabuľky relačnej databázy spoločne s adresármi *asset store* tvoria kľúčové položky pri plánovaní pravidelného zálohovania. Samotný obsah *asset store* nemá bez informácií z databázy pre používateľa hodnotu.
- Systém DSpace v mnohých smeroch výrazne spolieha na systémového administrátora, ktorý pomocou plánovača úloh (napr. *cron)* zabezpečí pravidelné vykonávanie dôležitých úloh a skriptov. Touto cestou je možné pravidelne prehliadať *asset store* kvôli zabezpečeniu integrity, extrakcii textu z voliteľných formátov a podobne.

Záver

Cieľom tohto príspevku nebolo určiť, ktorý zo skúmaných open-source systémov Invenio a DSpace je lepší či horší v oblasti dlhodobého uchovávania digitálnych objektov. Chceli sme poukázať na isté spoločné i rozdielne znaky, ktorými sú tieto systémy typické pri procesoch označovania a ukladania jednotlivých súčastí digitálnych objektov. Samotní používatelia mnohé zo spomínaných charakteristík ani nevnímajú, no pre digitálne objekty a ich dlhodobé bezpečné uchovávanie z pohľadu administrátora, ide o dôležité aspekty hodné osobitného zreteľa. Výsledky uskutočneného experimentu tak môžu pomôcť pri identifikácii potenciálne problémových oblastí a zraniteľností, ďalej pri zlepšovaní celkovej funkčnosti open-source systémov, ako aj pri plánovaní zálohovania a individuálnej príprave tzv. *disaster recovery* plánov, v prípade menších i väčších digitálnych repozitárov či digitálnych knižníc.

Veríme, že sme týmto príspevkom ponúkli odbornej verejnosti iný, ucelenejší pohľad na súčasti tvoriace digitálny objekt. Námetom pre budúce skúmanie tejto problematiky by mohli byť ďalšie testy spomínaných dátových štruktúr, či benchmarking správania sa systémov, napríklad, pri zaťažení vplyvom vysokého počtu uložených objektov.

Použitá literatúra

ANDRO, M. et al., 2012. Digital libraries: Comparison of 10 software. In: *Library Collections, Acquisitions, and Technical Services*, vol. 36, Issues 3-4, p. 79-83. ISSN 1464-9055

BARTOŠEK, M., 2004. Digitální knihovny – teorie a praxe. In: *Národní knihovna – knihovnická revue* [online], roč. 15, č. 4, s. 233 – 254 [cit. 2016-04-10]. Dostupné na: *http://full.nkp.cz/nkkr/NKKR0404/0404233.html*

CERN, 2015. *Invenio – common concepts* [online]. 2015 [cit. 2016-04-20]. Dostupné na:

http://invenio.readthedocs.org/en/latest/technology/common-concepts.html

SAIDIS, K.; A. DELIS, 2007. Type-consistent Digital Objects. In: *D-lib magazine* [online]. 2007, vol 13, 5/6 [cit. 2016-01-20]. ISSN 1082-9873. Dostupný na: *http://www.dlib.org/dlib/may07/saidis/05saidis.html*

THE DSPACE DEVELOPER TEAM, 2016. *DSpace 5.x documentation* [online]. 2016-03-21 [cit. 2016-04-11]. Dostupné na: *https://github.com/DSpace/DSpace/releases/download/dspace-5.5/DSpace-Manual.pdf*

UNIVERSITY OF CALIFORNIA, 2011. *Guidelines for Digital objects* [online]. 2011, 8 [cit. 2016-04-05]. Dostupné na: *http://www.cdlib.org/services/access_publishing/dsc/contribute/docs/GDO.pdf*

PhDr. Matúš Formanek

matus.formanek@mediamatika.sk ■

(Katedra mediamatiky a kultúrneho dedičstva, Fakulta humanitných vied, Žilinská univerzita v Žiline)