UNIVERZITNÁ KNIŽNICA V BRATISLAVE

CDA 2018

Trvalá udržateľnosť a perspektívy ďalšieho rozvoja LTP archívov

Zborník príspevkov z 3. medzinárodnej konferencie
o dlhodobej archivácii

univerzitná knižnica
v bratislave

Bratislava, 2018

Zostavila
Mgr. Katarína Tomková

Autori príspevkov
Milan Rakús, Juraj Strnisko
Zuzana Kvašová
Piotr Pałka, Tomasz Traczyk
Miklós Lendvay
Jiří Bernas
Petr Kukač
Zdeněk Vašek, Petr Cajthaml, Eliška Pavlásková
Zoltán Lux
Stanislav Dzúrik
Márton Németh, László Drótos
Andrej Bizík
Jaroslav Zeman

Obálka a grafický návrh
DOLIS GOEN, s.r.o., Bratislava
Zborník neprešiel jazykovou úpravou.

CIP SR

CDA 2018 : Trvalá udržateľnosť a perspektívy ďalšieho rozvoja LTP archívov: zborník príspevkov z 3. medzinárodnej konferencie o dlhodobej archivácii : Bratislava, 8. 11. 2018 / zost. Mgr. Katarína Tomková, 1. vyd. – Bratislava : Univerzitná knižnica v Bratislave, 2018

LTP archívy. Centrálny dátový archív. Dlhodobé dôveryhodné digitálne úložisko. Prevádzka LTP archívov. Udržateľnosť LTP archívov.

ISBN 978–80–89303–67–0
ISSN 2453-9309

Obsah

Úvod

Dlhodobé uchovávanie rozsiahlych digitálnych zbierok informačných prameňov pribudlo v ostatných rokoch do štandardného portfólia kompetencií pamäťových inštitúcií. Knižnice sa úspešne vyrovnávajú s revolučnými zmenami v informačných procesoch a technológiách, preberajú nové zodpovednosti a ponúkajú riešenia aktuálnych výziev. Univerzitná knižnica v Bratislave v rokoch 2012 – 2014 realizovala v rámci Operačného programu Informatizácia spoločnosti národný projekt Centrálny dátový archív (CDA) zameraný na dlhodobé uchovávanie kultúrneho obsahu. Štyri roky úspešnej prevádzky boli popri praktických skúsenostiach príležitosťou na hlbší prienik do problematiky digitálnej archivácie vrátane formulácie konkrétnych problémov a tém na riešenie. Tomu zodpovedalo aj zameranie našich vedeckých konferencií. Témou 1. ročníka medzinárodnej konferencie CDA 2016 (UKB, Bratislava, 10. 11. 2016) boli Formátové výzvy LTP. 2. ročník medzinárodnej konferencie CDA 2017 (UKB, Bratislava, 9. 11. 2017) bola zameraná na Výmenu skúseností z prevádzky a budovania LTP archívov. V roku 2019 vyvrcholí päťročný cyklus prevádzky a prvá fáza udržateľnosti projektu CDA. Požiadavka na trvalé zabezpečenie prevádzky a dlhodobo udržateľného rozvoja CDA vyplýva zo samotnej povahy projektu a hodnoty chráneného obsahu a stáva sa naliehavou témou strategického plánovania.

Aj preto sa program konferencie CDA 2018 upriamil na trvalú udržateľnosť a perspektívy rozvoja LTP archívov a poskytol priestor na prezentácie stavu a zámerov v danej oblasti vo všetkých krajinách V4 – v Čechách, Maďarsku, Poľsku a na Slovensku. Našim cieľom bolo prispieť k úrovni poznania v danej oblasti formou vybraných nosných príspevkov a tiež prostredníctvom diskusií a výmeny praktických aj teoretických poznatkov a názorov v medzinárodnom kontexte. Pozornosť, ktorú touto formou venujeme témam dlhodobej ochrany „digitálnych“ znalostí je systematickým príspevkom k napĺňaniu ambície Univerzitnej knižnice v Bratislave, najstaršej vedeckej knižnice na Slovensku, v oblasti vedeckej a výskumnej činnosti. Dnes možno konštatovať, že aj vďaka konferencii sa na pôde Univerzitnej knižnice v Bratislave založili dobré základy budúcej systematickej a dlhodobej spolupráce zainteresovaných expertov a inštitúcií.

Tretí ročník medzinárodnej konferencie CDA 2018: Trvalá udržateľnosť a perspektívy rozvoja LTP archívov sa uskutočnila dňa 8. 11. 2017 v Univerzitnej knižnici v Bratislave. Príspevky sú v zborníku zoradené v poradí podľa programu konferencie.

Konferencia s medzinárodnou účasťou sa už tradične konala v rámci Týždňa vedy a techniky na Slovensku 2018, ktorého zámerom je zlepšiť vnímanie vedy a techniky v povedomí celej spoločnosti, popularizovať a prezentovať ich, informovať verejnosť o poznatkoch vedy a techniky a o nutnosti podporovať vedu a techniku, ktoré sú základom hospodárskeho a spoločenského pokroku a prispievajú k riešeniu globálnych problémov a výziev. Je pre nás cťou, že Univerzitná knižnica v Bratislave mohla významnou mierou prispieť k úspešnému priebehu Týždňa vedy a techniky na Slovensku 2018 prostredníctvom pracovného stretnutia významných expertov z oblasti dlhodobého uchovávania digitálneho obsahu v podobe medzinárodnej vedeckej konferencie, ktorá sa za uplynulé tri roky už stala neodmysliteľnou platformou na výmenu skúseností v tejto oblasti.

V mene organizátorov konferencie

Ing. Silvia Stasselová, generálna riaditeľka UKB
Ing. Alojz Androvič, PhD., odborný garant konferencie

Stav a perspektívy rozvoja Centrálneho dátového archívu

Milan Rakús, Juraj Strnisko, Univerzitná knižnica v Bratislave, Bratislava, SR

Abstrakt
Centrálny dátový archív je výsledkom riešenia rovnomenného národného projektu číslo 8 Centrálny dátový archív, ktorý realizovala v rokoch 2011 – 2014 Univerzitná knižnica v Bratislave v rámci Operačného programu Informatizácia spoločnosti prioritnej osi 2: Rozvoj pamäťových a fondových inštitúcií a obnova ich národnej infraštruktúry. Centrálny dátový archív má za sebou takmer štyri roky prevádzky (2015 – 2018). Príspevok je zameraný na opis súčasného stavu budovania Centrálneho dátového archívu a možné perspektívy jeho ďalšieho rozvoja v nasledujúcich rokoch.

Abstract
The central data archive is the result of the national project number 8: Central Data Archive, which was implemented in 2011-2014 by the University Library in Bratislava. The project was solved within the framework of Operational Program Informatization on Priority Axis 2: Development of memory fund institutions and renewal of their national infrastructure. The central data archive has nearly 4 years of operation (2015-2018). This paper is aimed at describing the actual state of bulding the Central data archive. And possible perspectives of his further development in following years.

1 Úvod

Národný projekt Centrálny dátový archív (CDA) [1] realizovala Univerzitná knižnica v Bratislave (UKB) v rámci Operačného programu Informatizácia spoločnosti prioritnej osi 2: Rozvoj pamäťových a fondových inštitúcií a obnova ich národnej infraštruktúry (OPIS PO2) [2]. Projekt bol financovaný zo štrukturálnych fondov EÚ (ERDF/EFRR) a štátneho rozpočtu SR.

Výsledkom riešenia projektu bol CDA, vybudovaný ako dlhodobé dôveryhodné úložisko digitálneho obsahu. CDA bol implementovaný v súlade s ISO štandardom STN ISO 14721:2014 (OAIS) [3].

CDA je tvorený dvomi navzájom geograficky vzdialenými lokalitami. V Bratislave (UKB) je to lokalita CDA-A a v Martine (Slovenská národná knižnica) lokalita CDA-B. Obe lokality fungujú autonómne a každá z nich dokáže plnohodnotne zastúpiť funkciu druhej v prípade poruchy alebo odstávky. Okrem dvoch aktívnych lokalít disponuje CDA aj pasívnym skladom archivačných médií v lokalite CDA-C, ktorý sa nachádza v Bratislave (UKB). Uvedené riešenie garantuje vysokú bezpečnosť a dostupnosť uložených dát.

CDA má v súčasnom období za sebou takmer štyri roky prevádzky (2015 – 2018).

2 Súčasný stav budovania Centrálneho dátového archívu

Prevádzka CDA je na obdobie udržateľnosti (2015 – 2020) finančne zabezpečená v rámci dlhodobého plánu, ktorý sa každoročne aktualizuje v kontrakte UKB so zriaďovateľom. Zmluva o poskytovaní servisných služieb (SLA) (https://www.crz.gov.sk/index.php?ID=2288584&l=sk) je uzavretá do 29. 1. 2021.

Aktuálna Čiastková zmluva na poskytovanie služieb podpory NON IKT CDA (https://www.crz.gov.sk/index.php?ID=3647308&l=sk) je podpísaná do 31. 12. 2018. Plánované rozpočty na jednotlivé roky obdobia udržateľnosti sa darí plniť a sú dostatočné.

Hardvérové riešenie, softvérové riešenie, základné procesy a organizačné zabezpečenie CDA bolo dostatočne popísané v [4], [5] a [8].

Inštitúcie, ktoré majú uzavretú s CDA UKB Dohodu o zverení obsahu na dlhodobú archiváciu v systéme CDA alebo Predbežnú dohodu na dlhodobú archiváciu v systéme CDA (http://cda.kultury.sk/sk/podpisane_dohody_2015) tvoria Určené spoločenstvo CDA. Počet členov Určeného spoločenstva sa v roku 2018 nezmenil.

Pamäťové a fondové inštitúcie OPIS PO2

(1) Slovenská národná knižnica (SNK)
(2) Slovenský národný archív (SNA)

(3) Slovenská národná galéria (SNG)
(4) Múzeum Slovenského národného povstania (SNP)
(5) Pamiatkový úrad Slovenskej republiky (PUR)
(6) Slovenský filmový ústav (SFU)
(7) Štátna vedecká knižnica v Prešove (SVK)
(8) Národné osvetové centrum (NOC)
(9) Slovenský ľudový umelecký kolektív (SLK)
(10) Depozit digitálnych prameňov UKB (DDP)

Dopytovo orientované projekty OPIS PO2
(11) Kancelária Ústavného súdu SR (KUS)
(12) Vojenský historický ústav (VHA)
(13) Štátny geologický ústav Dionýza Štúra (GEO)
(14) Trnavský samosprávny kraj (TSK)
(15) Nitriansky samosprávny kraj (NSK)

Iné pamäťové a fondové inštitúcie
(16) Odbor digitalizácie UKB (UKB)
(17) Slovenská knižnica pre nevidiacich Mateja Hrebendu v Levoči (SKN)

CDA UKB sa stretáva s členmi Určeného spoločenstva na pracovných poradách podľa potreby. S Určeným spoločenstvom pravidelne komunikuje prostredníctvom cielených e-mailov o novinkách, podujatiach, ktoré organizuje, odstávkach systému a pod. Určené spoločenstvo má na webovej stránke CDA (http://cda.kultury.sk/) vyčlenenú sekciu s dokumentmi, prístupnú len pre členov komunity.

Štatistika vkladov do CDA k 30. 9. 2018 je uvedená na Obr. č. 1.

Kapacita CDA je naprojektovaná na 25 PB dát.

Štatistika výberov z CDA k 30. 9. 2018 je uvedená na Obr. č. 2.

Dopadový ukazovateľ Počet inštitúcií zapojených do vytvorených centier (6) je naplnený a prekročený (17). Dopadový ukazovateľ Počet novovytvorených pracovných miest (12) je naplnený. Z objektívnych dôvodov ponuky IT odborníkov na trhu práce sa nedarí dosiahnuť rovnaký počet mužov a žien na pracovisku.

IDT	Pamäťová a fondová inštitúcia (PFI)	Počet vložených SIP balíkov v CDA	Objem v TB
SNK	Slovenská národná knižnica	243 849	84,96
SNA	Slovenský národný archív	1 265 253	771,41
SNG	Slovenská národná galéria	10 184	14,81
SNP	Múzeum Slovenského národného povstania	154 882	2 740,3
PUR	Pamiatkový úrad Slovenskej republiky	196	6,289
SFU	Slovenský filmový ústav	1 110	664,55
SVK	Štátna vedecká knižnica v Prešove	0	0
NOC	Národné osvetové centrum	0	0
SLK	Slovenský ľudový umelecký kolektív	0	0
DDP	Depozit digitálnych prameňov UKB	4 895	28,213
KUS	Kancelária Ústavného súdu SR	0	0
VHA	Vojenský historický ústav	20 486	0,3245
GEO	Štátny geologický ústav Dionýza Štúra	16 560	1,51
TSK	Trnavský samosprávny kraj	0	0
NSK	Nitriansky samosprávny kraj	0	0
UKB	Odbor digitalizácie UKB	2 324	59,07
SKN	Slovenská knižnica pre nevidiacich Mateja Hrebendu v Levoči	3 862	15,182
	Spolu:	**1 723 601**	**4 386,6185**

Obr. č. 1. Štatistika vkladov do CDA k 30. 9. 2018

IDT	Pamäťová a fondová inštitúcia (PFI)	Počet diseminovaných DIP balíkov z CDA	Objem v TB
SNK	Slovenská národná knižnica	0	0
SNA	Slovenský národný archív	22	0,049
SNG	Slovenská národná galéria	0	0
SNP	Múzeum Slovenského národného povstania	78 028	1 424,995
PUR	Pamiatkový úrad Slovenskej republiky	0	0
SFU	Slovenský filmový ústav	12	0,19
SVK	Štátna vedecká knižnica v Prešove	0	0
NOC	Národné osvetové centrum	0	0
SLK	Slovenský ľudový umelecký kolektív	0	0
DDP	Depozit digitálnych prameňov UKB	0	0
KUS	Kancelária Ústavného súdu SR	0	0
VHA	Vojenský historický ústav	0	0
GEO	Štátny geologický ústav Dionýza Štúra	0	0
TSK	Trnavský samosprávny kraj	0	0
NSK	Nitriansky samosprávny kraj	0	0
UKB	Odbor digitalizácie UKB	23	0,4597
SKN	Slovenská knižnica pre nevidiacich Mateja Hrebendu v Levoči	0	0
	Spolu:	**78 089**	**1 425,702**

Obr. č. 2. Štatistika výberov z CDA k 30. 9. 2018

V oblasti práce s formátmi súborov boli v priebehu roku 2018 z kategórie perspektívnych formátov preradené do kategórie podporovaných formátov zaradené formáty PUID = fmt/16 (pdf 1.2), PUID = fmt/276 (pdf 1.7), PUID = fmt/289 (warc) a pribudli nové podporované formáty PUID = fmt/141 (wav) a PUID = fmt/143 (wav).

Formáty, ktoré akceptuje CDA k 30. 9. 2018 sú uvedené na Obr. č. 3.

Formáty, ktoré bude CDA výhľadovo akceptovať v budúcnosti sú uvedené na Obr. č. 4.

PUID	MIME TYPE	Identifikátor	Validátor	Poznámka
cda/101	application/vnd.cda.container.x-dpx	Proprietárny pre CDA	JHOVE (modul BYTESTREAM)	Proprietárny PUID pre CDA[1]
cda/102	application/vnd.cda.container.pusr	Proprietárny pre CDA	JHOVE (modul BYTESTREAM)	Proprietárny PUID pre CDA[2]
fmt/1	audio/x-wav	DROID	JHOVE	
fmt/2	audio/x-wav	DROID	JHOVE	
fmt/6	audio/x-wav	DROID	JHOVE	
fmt/16	application/pdf	DROID	JHOVE	PDF 1.2
fmt/17	application/pdf	DROID	JHOVE	PDF 1.3
fmt/18	application/pdf	DROID	JHOVE	PDF 1.4
fmt/19	application/pdf	DROID	JHOVE	PDF 1.5
fmt/20	application/pdf	DROID	JHOVE	PDF 1.6
fmt/41	image/jpeg	DROID	JHOVE	
fmt/42	image/jpeg	DROID	JHOVE	JPEG 1.00
fmt/43	image/jpeg	DROID	JHOVE	JPEG 1.01
fmt/44	image/jpeg	DROID	JHOVE	JPEG 1.02
fmt/94	model/vrml	DROID	Chisel	Starý validátor
fmt/101	text/xml	DROID	JHOVE	
fmt/141	audio/x-wav	DROID	JHOVE 1.18	
fmt/142	audio/x-wav	DROID	JHOVE	
fmt/143	audio/x-wav	DROID	JHOVE 1.18	
fmt/156	image/tiff	DROID	JHOVE	
fmt/193	application/octet-stream	DROID	JHOVE (modul BYTESTREAM)	DPX 1.0
fmt/276	application/pdf	DROID	JHOVE 1.18	PDF 1.7
fmt/289	application/warc	DROID	warctools	Web Archive
fmt/353	image/tiff	DROID	JHOVE	
fmt/355	application/rtf	DROID	JHOVE (modul BYTESTREAM)	
fmt/436	image/tiff	DROID	JHOVE	
fmt/541	application/octet-stream	DROID	JHOVE (modul BYTESTREAM)	DPX 2.0
fmt/645	image/jpeg	DROID	JHOVE	
fmt/703	audio/x-wav	DROID	JHOVE	
fmt/704	audio/x-wav	DROID	JHOVE	
x-fmt/111	text/plain	Enca	JHOVE (modul UTF8-hul)	UTF-8 bez BOM
x-fmt/387	image/tiff	DROID	JHOVE	
x-fmt/391	image/jpeg	DROID	JHOVE	
x-fmt/392	image/jp2	DROID	JHOVE	

Poznámky:

1. cda/101 je kontajner pre Slovenský filmový ústav (SFÚ)
2. cda/102 je kontajner pre Pamiatkový úrad SR (PÚ SR)

Obr. č. 3. Formáty, ktoré akceptuje CDA k 30. 9. 2018

PUID	MIME TYPE	Identifikátor	Validátor	Poznámka
cda/103	application/x.cda.noc-xdpx	Proprietárny pre CDA	JHOVE (modul BYTESTREAM)	Proprietárny PUID v CDA[1]
fmt/5	video/x-msvideo	DROID	MediaConch	Kontajner AVI, kodek FFV1 pre video, PCM pre audio
fmt/11	image/png	DROID	JHOVE 1.18 alebo novší / pngcheck	
fmt/12	image/png	DROID	JHOVE 1.18 alebo novší / pngcheck	
fmt/13	image/png	DROID	JHOVE 1.18 alebo novší / pngcheck	
fmt/14	application/pdf	DROID	JHOVE	PDF 1.0
fmt/15	application/pdf	DROID	JHOVE	PDF 1.1
fmt/95	application/pdf	DROID	veraPDF	PDF/A (1a)
fmt/354	application/pdf	DROID	veraPDF	PDF/A (1b)
fmt/476	application/pdf	DROID	veraPDF	PDF/A (2a)
fmt/477	application/pdf	DROID	veraPDF	PDF/A (2b)
fmt/483	application/epub+zip	DROID	epubcheck	
fmt/569	video/x-matroska	DROID	MediaConch	Kontajner Matroska, kodek FFV1 pre video, PCM pre audio

Poznámky:

1. cda/103 je kontajner pre Národné osvetové centrum (NOC)

Obr. č. x: Formáty, ktoré bude CDA výhľadovo akceptovať v budúcnosti

Obr. č. 4. Formáty, ktoré bude CDA výhľadovo akceptovať v budúcnosti

Pripomíname, že v súčasnom období sa na opis obsahu súborov čoraz častejšie používa identifikátor media type (MIME type, content type). Hodnoty media type celosvetovo definuje (prideľuje, eviduje) Internet Assigned Numbers Authority (IANA) Autorita pre prideľovanie čísel na Internete (http://www.iana.org/). Na Obr. č. 5 sú uvedené počty mime type evidované v registroch IANA (http://www.iana.org/assignments/media-types/media-types.xhtml) v dňoch 27. 9. 2016, 1. 10. 2017 a 4. 10. 2018.

TOP LEVEL TYPE	Stav k 27. 9. 2016	Stav k 1. 10. 2017	Stav k 4. 10. 2018
application	1190	1253	1314
audio	144	148	150
font	0	6	6
example	1	1	1
image	56	56	63
message	21	21	21
model	22	23	26
multipart	15	17	17
text	72	73	75
video	78	79	81
Spolu:	**1599**	**1677**	**1754**

Obr. č. 4. Počty mime type evidované v registroch IANA

Organizačné zabezpečenie CDA UKB bolo dostatočne popísané v [5] a počas obdobia udržateľnosti je nemenné. Prevádzku a rozvoj CDA zabezpečuje 12 zamestnancov UKB. Podarilo sa stabilizovať pracovný kolektív. Počas celého obdobia prevádzky CDA sa traja zamestnanci vymenili, dve pracovníčky odišli na materskú dovolenku. Získavanie kvalifikovanej náhrady je veľmi problematické.

V rámci obnovy IKT CDA sme v súlade s rozpočtom na obdobie udržateľnosti projektu realizovali tieto aktivity:

- Obnova IKT CDA 2015
- Obnova IKT CDA 2016, I. etapa
- Obnova IKT CDA 2016, II. etapa
- Obnova IKT CDA 2017

V roku 2018 MK SR bez uvedenia dôvodu obmedzilo investičné prostriedky na obnovu IKT CDA.

V rámci bitovej ochrany (kontroly integrity) dlhodobo uchovávaných dát sme po viacerých testovacích obdobiach pristúpili od augusta 2018 k systematickej kontrole dát v lokalite CDA-B, od začiatku novembra 2018 pristúpime k systematickej kontrole dát v lokalite CDA-A. Kontrola integrity dát uložených v lokalite CDA-C je v procese riešenia. Pri bitovej ochrane dát sa kontrolujú dáta uložené na jednotlivých páskach. Kvôli predstave uvádzame, že v lokalite CDA-B boli v auguste 2018 skontrolované 4

archívne magnetické pásky na ktorých je uložených 10892 AIP (14,89718 TiB), v mesiaci september 2018 bolo skontrolovaných 16 archívnych magnetických pások na ktorých je uložených 54619 AIP (43,31678 TiB).

CDA UKB bol v období rokov 2014 – 2016 certifikovaný v súlade s normou „STN ISO/IEC 27001:2013 (SMIB) [6]. V roku 2014 sa uskutočnil certifikačný audit, v rokoch 2015 a 2016 sa realizoval dozorný audit. V roku 2017 sa uskutočnil recertifikačný audit, koncom roka 2018 plánujeme dozorný audit.

Do konca roku 2018 sa CDA pokúsi získať certifikát CoreTrustSeal. Certifikát CoreTrustSeal poskytuje organizácia DANS – Data Archiving and Networked Services z Holandska (https://www.knaw.nl/en/institutes/dans). Administratívny poplatok za udelenie certifikácie je 1 000 EUR. Certifikát sa udeľuje na dva roky a má veľmi silné postavenie najmä medzi európskymi inštitúciami a konzorciami (CESSDA), ktoré participujú na významných európskych projektoch (CLARIN, DARIAH, EUDAT).

Medzinárodné kolokvium, Brno 31. 5. – 1. 6. 2016, Knižnice krajín V4 v digitálnom veku, poskytlo rámec na spoluprácu knižníc v rámci rôznych platforiem.

V roku 2016 sme neformálne koncipovali LTP platformu krajín V4. V rámci LTP platformy krajín V4 sa uskutočnilo niekoľko pracovných stretnutí vybratých zástupcov zainteresovaných krajín. Prvé pracovné stretnutie sa uskutočnilo dňa 11. 11. 2016 v Univerzitnej knižnici v Bratislave (Slovensko, Česko, Maďarsko), druhé v dňoch 20. a 21. 6. 2017 v Národní knihovne ČR v Prahe (Česko, Slovensko), tretie 10. 11. 2017 v Univerzitnej knižnici v Bratislave (Česko, Slovensko), štvrté v dňoch 20. a 21. 6. 2018 v Národní knihovne ČR v Prahe (Česko, Slovensko). Pripravujeme piate stretnutie dňa 9. 11. 2018 v Univerzitnej knižnici v Bratislave (Slovensko, Česko, Maďarsko, Poľsko).

V dňoch 4. a 5. októbra 2018 sme prijali pozvanie na pracovné stretnutie odborníkov národných knižníc z juhovýchodnej Európy (SEENL framework; krajiny bývalej Juhoslávie, Bulharsko, Grécko a Albánsko). Väčšina z týchto krajín má zvládnuté procesy digitalizácie a sprístupňovania zdigitalizovaného obsahu napr. formou digitálnej knižnice (napr. Srbská digitálna knižnica dostupná na http://www.digitalna.nb.rs/), no s budovaním dlhodobých dôveryhodných dátových úložísk nemajú takmer žiadne skúsenosti. Vo väčšine prípadov sú len na začiatku cesty. Hľadajú možné technické riešenia, modely ich fungovania a financovania. Naše skúsenosti s budovaním a takmer štvorročnou prevádzkou Centrálneho dátového archívu Univerzitnej knižnice v Bratislave mali preto pre nich nesmiernu hodnotu.

V rámci interného vedecko-výskumného procesu zamestnanci CDA v rokoch 2016 a 2017 riešili vedecko-výskumného projekt UKB UAI-CDA-01: Rozvoj metód dlhodobej ochrany digitálnych prameňov s reálnymi výstupmi „SW produkt na tvorbu SIP balíkov s možnosťou validácie vstupných súborov EXCELMETS“ a metodický a štandardizačný materiál „Formáty súborov akceptované CDA UKB“. Na obdobie rokov 2018 a 2019 sme v rámci vedecko-výskumného projektu UKB UAI-CDA-01pripravili na riešenie ďalšie dve témy: „Optimalizácia diseminačných procesov v podmienkach Centrálneho dátového archívu“ a „Optimálny model práce s treťou archívnou kópiu dát“.

Progresívny charakter mali aj 1. medzinárodná konferencia CDA 2016: Formátové výzvy LTP, ktorá sa uskutočnila 10. 11. 2016 v UKB a 2. medzinárodná konferencia CDA 2017: Výmena skúseností z prevádzky a budovania LTP archívov, ktorá sa konala, 9. 11. 2017, takisto v UKB. Dnes, 8. 11. 2018, sa koná konferencia CDA 2018: Trvalá udržateľnosť a perspektívy ďalšieho rozvoja LTP archívov.

Všetky výsledky práce CDA sú zverejňované na webovej stránke http://cda.kultury.sk/. Stránku pravidelne aktualizujú zamestnanci CDA s využitím redakčného systému Drupal. Bohatá publikačná činnosť zamestnancov CDA je zverejňovaná v Správach o činnosti a hospodárení UKB za jednotlivé roky. Nezanedbateľné sú aj vystúpenia zamestnancov CDA na národných a medzinárodných konferenciách a podujatiach podobného charakteru. Časté sú aj exkurzie v CDA.

3 Perspektívy rozvoja Centrálneho dátového archívu

Centrálny dátový archív bol projektovaný ako LTP (Long Term Preservation) archív (dlhodobé dôveryhodné úložisko). Pri takomto type archívu sa predpokladá, že informácie budú v ňom uložené veľmi dlho, mali by byť stále čitateľné a mali by byť neustále prístupné používateľovi. Prevádzka archívu nie je jednoduchá. Informačné a komunikačné technológie sa vyvíjajú obrovskou rýchlosťou. Hardvér a softvér zastaráva, formáty súborov sú poznačené prudkými zmenami (vznik nových formátov, vývoj existujúcich formátov, postupné zanikanie nepodporovaných formátov). LTP archívy musia eliminovať hrozby a riziká spojené s dlhodobým uchovávaním digitálneho obsahu. Musia byť zabezpečené proti strate dát. Musia byť odolné proti vonkajším a vnútorným útokom, musia neustále obnovovať HW a SW, musia byť dlhodobo finančne a personálne zabezpečené, musia byť transparentné a pod. [5].

V súčasnom období nie je jasné akým smerom sa budú uberať projekty riešené v rámci OPIS PO2 [2] po skončení obdobia udržateľnosti. CDA UKB má pri ochrane kultúrneho dedičstva svoje dlhodobé opodstatnenie.

3.1 Krátkodobý výhľad rozvoja CDA

Krátkodobým výhľadom rozvoja CDA rozumieme výhľad do konca obdobia udržateľnosti projektu.

V rámci Krátkodobého výhľadu rozvoja CDA uvažujeme napr.:

- Zorganizovať medzinárodnú konferenciu CDA 2019: Nové trendy v budovaní LTP archívov
- Certifikovať CDA podľa normy STN ISO/IEC 27001:2013 (SMIB) [6] (V priebehu rokov 2017 – 2019)
- Certifikovať CDA ako LTP archív podľa normy CoreTrustSeal – Osvedčenie dôveryhodnosti
- Nahradiť hlavnú páskovú knižnicu (nosná časť CDA) knižnicou novšej generácie

3.2 Dlhodobý výhľad rozvoja CDA

Dlhodobým výhľadom rozvoja rozumieme výhľad po skončení obdobia udržateľnosti projektu na najbližších päť rokov.

V rámci Dlhodobého výhľadu rozvoja CDA uvažujeme napr.:

- Optimalizovať HW, SW a organizačné zabezpečenie CDA v súlade s novými trendmi v tejto oblasti
- Znížiť náklady spojené s prevádzku CDA UKB
- Poskytovať služby Určenému spoločenstvu nad rámec projektu (tvorba SIP balíkov, formátové konverzie a pod.)
- Poskytnúť kapacity CDA ďalším organizáciám

CDA sa bude aj naďalej v pravidelných intervaloch zaoberať identifikáciou a riešením formátových rizík. Základným nástrojom na identifikáciu a riešenie formátových rizík je Formátová databáza CDA [5]. V pravidelných (mesačných) intervaloch sa synchronizuje s databázou PRONOM. Rozdiely slúžia ako podklad pre rozhodovanie o prípadnej formátovej konverzii alebo o iných opatreniach.

CDA bude aj naďalej sledovať vývoj v oblasti formátov a nástrojov na identifikáciu, validáciu a konverziu obsahu súborov.

CDA sa bude aj naďalej zapájať do projektov, ktoré zabezpečujú vývoj, testovanie a využívanie nástrojov na identifikáciu, validáciu a konverziu obsahu súborov, ako to bolo napr. pri projekte PREFORMA (http://www.preforma-project.eu/index.html).

4 Záver

Realizácia projektu CDA na dlhodobú archiváciu digitalizovaných kultúrnych objektov korešponduje s dlhodobou stratégiou rozvoja slovenského knihovníctva a odpovedá poslaniu UKB v roli novodobej pamäťovej inštitúcie. Prax ukázala, že úsilie na vytvorenie podmienok na dôveryhodnú a spoľahlivú ochranu digitalizovaného kultúrneho dedičstva je kontinuálny proces optimalizácie a inovácie technických, metodických a organizačných aspektov špecifického informačného systému. Dôvodom je jednak neustály rozvoj a zmeny v prostredí tvorby a využívania digitálnych dokumentov, ako aj doterajšie skúsenosti nadobudnuté počas prvých rokov prevádzky CDA. Nevyčísliteľnú hodnotu má nielen archivovaný obsah ale aj získané know-how, ktoré reflektuje aktuálny vývoj vo svete a má trvalý výskumný a vývojový potenciál. Do tohto rámca spadajú aj certifikačné aktivity CDA, ktoré predstavujú cyklickú verifikáciu vlastností, podmienok a zručností pri realizácii a zabezpečení prevádzky systémov LTP. To všetko sa deje s cieľom dlhodobo udržateľného rozvoja digitálnych služieb UKB v intenciách zadania zriaďovateľa, Ministerstva kultúry Slovenskej republiky.

Centrálny dátový archív je jediným archívom tohto typu na Slovensku. Je jedným z prvých úspešných projektov podobného rozsahu v Európe. Pevne veríme, že vláda SR zabezpečí financovanie systému CDA v dlhodobom časovom horizonte a MK SR bude nastúpenému trendu CDA venovať primeranú pozornosť.

Na úplný záver chceme poďakovať svojim kolegom Ing. Stanislavovi Lichému a Mgr. Kataríne Tomkovej za pomoc s prípravou vecných podkladov pre tento príspevok.

Použité skratky

AIP – Archival Information Package (Archívny informačný balík)
CDA – Centrálny dátový archív
HW – Hardvér
OPIS – Operačný program Informatizácia spoločnosti
PFI – Pamäťová a fondová inštitúcia
PO – Prioritná os
PUID – Persistent Unique Identifier (Unikátny identifikátor formátu registrovaný službou PRONOM)
SIP – Submission Information Package (Vkladaný informačný balík)
SLA – Service Level Agreement
SW – Softvér
SMIB – Systém manažérstva informačnej bezpečnosti
UKB – Univerzitná knižnica v Bratislave

Použitá literatúra

[1] CIGLAN, Ivan: Národný projekt Centrálny dátový archív. In: ITlib, 2.2012, s. 35-36

[2] Operačný program Informatizácia spoločnosti – Prioritná os 2

[3] STN ISO 14721:2014: Systémy prenosu vesmírnych údajov a informácií. Otvorený archívny informačný systém (OAIS). Referenčný model

[4] ANDROVIČ, Alojz et al.: Centrálny dátový archív v roku 1. In: ITlib, 2.2016, s. 37-52

[5] RAKÚS, Milan. CDA a Formátová stratégia CDA. In: *CDA 2016 Formátové výzvy LTP: zborník príspevkov z 1. medzinárodnej konferencie o dlhodobej archivácii.* Bratislava: Univerzitná knižnica v Bratislave, 2016, s. 7-19. ISSN 2453-9406.

[6] STN ISO/IEC 27001:2013 Systém manažérstva informačnej bezpečnosti (SMIB)

[7] STN ISO 16363:2014 : Systémy prenosu vesmírnych údajov a informácií. Audit a certifikácia dôveryhodných digitálnych úložísk

[8] RAKÚS, Milan. Tri roky prevádzky Centrálneho dátového archívu. In: *CDA 2017 Výmena skúseností z prevádzky a budovania LTP archívov: Zborník príspevkov z 2. medzinárodnej konferencie o dlhodobej archivácii.* Bratislava: Univerzitná knižnica v Bratislave, 2017, s. 7-21. ISSN 2453-9309.

Stav dlouhodobé archivace v NK ČR

Zuzana Kvašová, Národní knihovna České republiky, Praha, ČR

Abstrakt:
Příspěvek se týká otázek dlouhodobé archivace digitálních dokumentů v NK ČR, aktuálního stavu standardizace a procesů ukládání do dlouhodobého úložiště. NK ČR pokračuje ve vydávání standardů pro nové typy dokumentů s přesahem samotné digitalizace probíhající v této instituci a pro potřeby praktické implementace doporučuje archivační formáty a validační nástroje, vytváří metodiky a vyvíjí vlastní komplexní validační nástroj. Příspěvek rovněž představí Koncepci rozvoje knihoven v oblasti dlouhodobé archivace, zejména s ohledem na legislativu a úpravy statutárních dokumentů.

Abstract:
The paper addresses issues of long-term archiving of digital documents in the NL CR, the current state of standardization and preservation processes in the long-term repository of NL CR. The NL CR continues to publish standards for new types of documents outside the scope of in-house digitization. For practical implementation needs NL CR recommends archival formats and validation tools, creates methodologies and develops in-house complex validation tool. The paper will also introduce The Strategy for the Development of Libraries for the long-term archiving, especially with regard to legislation and charter documents updating.

Úvod

Digitalizace různých typů dokumentů je v knihovnách i dalších kulturních institucích v České republice již zavedený proces, který má podporu mnoha finančních mechanismů, je na něj myšleno v koncepčních materiálech. Digitalizace jako jeden ze způsobů zachování kulturního dědictví se tedy již stalo nezpochybnitelnou a ustálenou součástí agend těchto institucí a řada z nich ho zařadila mezi své běžné činnosti. Digitalizace je pojímána především jako způsob, jak kulturní objekty zpřístupnit širšímu spek-

tru uživatelů. Ať se již na digitalizaci díváme tímto způsobem, nebo na ni hledíme i jako na způsob ochrany kulturního dědictví (což je další ze zásadních a nezpochybnitelných aspektů), je třeba finance vynaložené na tuto činnost zhodnotit tím, že zajistíme dlouhodobou udržitelnost digitálních dat, které při ní vznikají.

Přestože téma digitální archivace není mezi kulturními institucemi nic nového, řada institucí tuto oblast neřeší, nebo neví, jak řešit. V rámci České republiky byl jedním z prvních projektů, který tuto tematiku zahrnul do svých cílů, projekt Vytvoření národní digitální knihovny. Projekt financovaný z větší části evropským dotačním mechanismem IOP, běžel v Národní knihovně v letech 2010 – 2014, od roku 2015 se nachází ve fázi udržitelnosti. V tomto rozsahu na něj zatím v oblasti knihovnictví žádný další projekt nenavázal. Digitalizační činnost i digitální archivace jsou již řadu let zahrnovány do různých strategických materiálů na národní úrovni, odkud se poté propojeny do dalších koncepčních dokumentů nižší úrovně.

Podpora dlouhodobé archivace v ČR

Klíčovými dokumenty na národní úrovni jsou v této oblasti především *Státní kulturní politika*[1] a *Strategie Evropa 2020 – Digitalizace kulturního obsahu*[2](dále jako *Strategie digitalizace kulturního obsahu*). Oba dokumenty byly vytvořeny s výhledem do roku 2020, Státní kulturní politika je pak pravidelně aktualizována, momentálně počítá s výhledem k roku 2025.

Dlouhodobá archivace je v těchto klíčových dokumentech zahrnuta, pojata jako nezbytná součást péče o digitalizované kulturní dědictví s napojením na zpřístupnění objektů a poskytnutím nástrojů pro instituce, které se digitalizací zabývají.

Státní kulturní politika je zásadním koncepčním materiálem, který je pravidelně aktualizován a vyhodnocován. Významnou součástí dokumentu z hlediska dlouhodobé archivace je především priorita *2. Rozvoj kreativity, podpora kulturních činností a vzniku kulturních statků, poskytování veřejných kulturních služeb, práce s publikem, podpora přístupu ke kultuře a rozvoj participativní kultury usnadňující sociální začlenění* a priorita *3. Uchování kulturního dědictví*. Konkrétně opatření *2.2.3 Pokračovat ve vybavování knihoven a dalších paměťových institucí potřebnými technologiemi a informačními zdroji, zajistit dlouhodobou udržitelnost jejich provozuschopnosti, optimalizovat systém, vytvořit kompetenční centrum na MK, vybudovat robustní ICT in-*

1 https://www.mkcr.cz/statni-kulturni-politika-69.html

2 https://www.mkcr.cz/strategie-evropa-2020-digitalizace-kulturniho-obsahu-831.html

frastrukturu v gesci MK včetně LTP archivu a opatření *3.3.4 Podporovat důvěryhodné dlouhodobé uložení digitálních dokumentů*. V důvodové zprávě je předložen rozvoj ICT v přímé gesci Ministerstva kultury ČR, dokument předpokládá vybudování robustní ICT infrastruktury včetně LTP archivu – s dosahem napříč resortem kultury, ale i mimo něj. Na druhou stranu se v dokumentu mluví o více depozitářích zajišťujících uchování digitalizovaného kulturního dědictví. Jako vysoce žádoucí je deklarována potřeba certifikace depozitářů pro dlouhodobé uchování dokumentů. Akcentována je rovněž potřeba zajištění nástrojů pro podporu dlouhodobého uložení digitálních dokumentů. (1)

Strategie digitalizace kulturního obsahu je dokument platný pro léta 2013 – 2020. S tím, že součástí je deklarace nutné aktualizace a rozšíření. Zásadními cíli pro oblast dlouhodobé archivace jsou strategické cíle *3. Bezpečné uchování digitálních dokumentů* a *4. Vytvoření organizačních a technických předpokladů trvalého uchování a zpřístupnění digitálních dokumentů včetně ustavení zvláštní pracovní skupiny*, plus dále cíl *6. Zajištění finančních prostředků*, který s předchozími body úzce souvisí. Základní strategické cíle tohoto dokumentu jsou dále rozpracovány v příloze, kde jsou specifikovány dle jednotlivých oblastí resortu kultury. Pro oblast knihoven je v dokumentu deklarován vznik dlouhodobého důvěryhodného úložiště na národní úrovni, kromě toho také zavedení persistentních identifikátorů digitálních objektů na národní úrovni a vývoj finálního mechanismu institutu povinného ukládání dokumentů. Strategie dále předpokládá úzkou spolupráci knihoven v této oblasti, udržování společných nástrojů typu Registr digitalizace, Systém ČIDLO[3] a v neposlední řadě legislativní podporu těchto činností především formou zajištění povinného výtisku elektronických publikací, řešení děl nedostupných na trhu a osiřelých děl. (2)

Řada těchto cílů je již splněna, nebo alespoň částečně směřují k naplnění. Část cílů se naplnit nepodařilo. Především systém pro ukládání digitálních dat – vybudování úložiště na národní úrovni (v dokumentu konkrétně bod 2. Vybudování důvěryhodného úložiště digitálních dokumentů na národní úrovni v rámci cíle *3. Bezpečné uchování digitálních dokumentů*), který byl zamýšlen vytvořit v Národní knihovně ČR.

V rámci setkávání skupiny tzv. Sektorových agregátorů (tj. zástupci jednotlivých příspěvkových organizací napříč resortem) pod Ministerstvem kultury ČR bylo v roce 2016 provedeno vyhodnocení cílů tohoto dokumentu za účelem jeho aktualizace. Snahou bylo vytvořit strategii zaměřenou na digitalizaci, uchovávání (dlouhodobou archivaci) a zpřístupnění digitálních dokumentů s ohledem na současné strategické dokumenty a existující dotační mechanismy. Tu se bohužel dosud nepodařilo realizovat.

3 https://resolver.nkp.cz

Oba zásadní dokumenty se ve velké míře shodují a doplňují, problematickým tématem se zdá především pojetí realizace těchto snah a následným poskytnutím vhodného systému pro dlouhodobé ukládání digitálních dokumentů na resortní úrovni.

Koncepce dlouhodobé archivace v knihovnách

V prostředí knihoven je na výše uvedené materiály navázáno především dokumentem *„Národní koncepce dlouhodobé ochrany digitálních dat v knihovnách“* (3), který vznikl v rámci implementace *Koncepce rozvoje knihoven ČR na léta 2011–2015*. (4) V roce 2016 byl aktualizován s ohledem na nově připravovanou *Koncepci rozvoje knihoven na léta 2017 – 2020* (3), v prosinci 2016 byl poté schválen Ústřední knihovnickou radou. Účelem koncepce je připravit v národním měřítku podmínky pro uchovávání a dlouhodobou ochranu digitální dat v knihovnách, s ohledem na jejich roli klíčových informačních institucí. Dokument definuje kroky, které je vhodné učinit, aby byla dlouhodobá ochrana účinná. Vychází z existence Standardu NDK a systému pro trvalou identifikaci (URN:NBN), které jsou využívány pro většinu digitalizačních projektů zaměřených na novodobé fondy. Jednotný standard, který existuje jak pro novodobé, tak historické dokumenty, je silná stránka současného stavu. Standard je nutné vyvíjet kvůli novým typům dokumentů a vývoji poznání v oblasti dlouhodobého uchovávání digitálních dokumentů. Koncepce předpokládá několik zásadních procesů, které by se měly postupně prolínat. Klíčovým krokem pro dlouhodobou archivaci je prosazení péče o digitální data do trvalých aktivit institucí, její organizační zakotvení (jak úroveň jednotlivých knihoven, tak v celé knihovní síti) a formální provádění (vytváření nutné dokumentace). Oproti výše uvedeným koncepčním dokumentům na národní úrovni je očekávána diverzifikace dlouhodobé ochrany mezi více institucí a tím i více přístupů. Přínos je očekáván v kooperaci mezi jednotlivými knihovnami i mimo knihovní sektor (především s akademickými institucemi). Důležitým principem, který koncepce vyzdvihuje, je nízkoprahovost přístupu k informacím, technologiím a službám spojeným s péčí o digitální data – tedy aby každý, kdo tato data vlastní, měl přístup k možnostem jejich ochrany. Stejně jako u dalších materiálů i zde je zmíněna nutnost zajištění institucionálního financování dlouhodobého uchovávání digitálních dat z veřejných rozpočtů.

Jako zásadní je v této koncepci uvedeno zřízení Metodického centra pro dlouhodobou ochranu digitálních dat, které by mělo vzniknout při Národní knihovně ČR. Dokument je podpořen i vedením Národní knihovny, o finance na Metodické centrum bylo žádáno Ministerstvo kultury v uplynulých letech.

Samotná problematika realizace dlouhodobé archivace je v dokumentu řešena návrhem sítě úložišť, kdy by vedle garantovaného datového centra (pro ochranu bit-stream) a státem garantovaného dlouhodobého úložiště (navazuje na obě výše uvedené koncepce), měla vzniknout lokální úložiště, ideálně založená na odlišných technologických řešeních. Otázkou je, jak se tyto cíle podaří realizovat. Současný směr se jeví spíše v rozšiřování kapacity dlouhodobého úložiště v NK ČR, s doplněním o další úložiště pro jednotlivé instituce s různým stupněm dostupných procesů dlouhodobé archivace.

Realizace dlouhodobé archivace v knihovnách

V současné době je v Národní knihovně ČR provozováno dlouhodobé úložiště, které bylo vytvořeno v rámci projektu Národní digitální knihovny. Úložiště je dostupné pro data, která jsou vytvořena produkčními linkami partnerských institucí projektu – Národní knihovnou a Moravskou zemskou knihovnou. Přijímána jsou navíc data, která vznikají podle stejného využívaného standardu v rámci dotačního programu VISK 7. Řada knihoven dlouhodobě žádá o možnost uložit v tomto systému svá data (navíc využívají standard, který k tomuto účelu byl vytvořen, tzv. Standard NDK[4]) (5). To dosud nebylo umožněno, především z infrastrukturních a finančních důvodů. Pro tento krok je třeba, aby byla zabezpečena dlouhodobá podpora úložiště (především ze strany zřizovatele), v ideálním případě by úložiště mělo být certifikováno.

V roce 2016 byla zahájena realizace projektu ArcLib[5] ve spolupráci několika partnerských institucí – Knihovny Akademie věd, Masarykovy Univerzity, Národní knihovny ČR a Moravské zemské knihovny. Cílem projektu je vytvořit řešení pro dlouhodobou archivaci digitálních dokumentů založeného na open-source nástrojích, pro užití knihovnami v ČR. Výstupy projektu by měly být dostupné po ukončení projektu v letech 2020. Je předpokládána kompatibilita s řešením provozovaným v Národní knihovně ČR.

V roce 2018 rovněž spustilo sdružení CESNET službu dlouhodobého uložení dat (bez plného procesu LTP), kterou nabídlo i knihovnám v ČR. Tuto službu však podle aktuálních informací zatím žádná knihovna nevyužívá.

Ačkoliv je tedy problematika dlouhodobé archivace součástí koncepčních materiálů, realita stále ukazuje, že digitální data vytvářená knihovnami a dalšími institucemi

4 https://www.ndk.cz/standardy-digitalizace/standardy-digitalizace-1

5 https://arclib.cz/

v ČR z větší části nejsou dlouhodobě uchovávána. Příčin tohoto stavu je vícero. Jedním z důvodů je vyplývající nekonzistence ve způsobu řešení provozu dlouhodobého úložiště (úložišť). Starší materiály počítaly se vznikem jednoho centrálního úložiště, které měla podle původních představ provozovat Národní knihovna ČR. Ze státní kulturní politiky vyplývá představa resortního úložiště provozovaného přímo v gesci MK ČR. Naproti tomu aktuální koncepce knihoven počítá se sítí úložišť, která se budou vzájemně doplňovat. Ať už bude přikročeno k jakékoliv variantě, zásadním faktorem je dostatečné financování pro zřízení a především pro provoz. V současné době nemá žádná instituce (mimo NK ČR) zajištěno financování dlouhodobé archivace, v NK ČR navíc chybí výhled do budoucna.

Současné aktivity knihoven v oblasti dlouhodobé archivace

Od vydání nové Koncepce rozvoje knihoven na léta 2017-2020 (3) začaly pod Ústřední knihovnickou radou fungovat pracovní skupiny, z nichž jedna je přímo zaměřena na oblast digitální archivace (konkrétně pracovní skupina vzniklá v rámci bodu *1.4 Zabezpečit důvěryhodné, dlouhodobé uchování digitálních dokumentů*) a další se toho tématu dotýkají. Pracovní skupina, složená z odborníků napříč institucemi, řeší problematiku v návaznosti na výše zmíněnou *„Národní koncepci dlouhodobé ochrany digitálních dat v knihovnách"* (3), především v oblasti zajištění existence úložišť a standardizace. Koncepce předpokládá vznik Metodického centra pro dlouhodobou ochranu digitálních dat. Toto metodické centrum by mělo vzniknout v Národní knihovně ČR, která pro jeho vznik již podnikla několik kroků, mimo jiné požádala i o financování svého zřizovatele. Pracovní skupina také navrhla nový podprogram v rámci dotačního programu VISK[6], jehož cílem by mělo být umožnit knihovnám čerpat finance na procesy vedoucí k zajištění dlouhodobé digitální archivace. V neposlední řadě se pracovní skupina snaží zajistit dostatečnou legislativu pro archivaci digitálních dat a zajištění dlouhodobé ochrany. Výsledkem je návrh změny tzv. Knihovního zákona, který by měl obsahovat povinnost knihoven zajistit trvalou ochranu svých digitálních dat.

Národní knihovna v souvislosti s činností pracovní skupiny a přípravou Metodického centra změnila fungování Formátového výboru[7]. Ten do roku 2018 fungoval jako poradní uskupení, složené ze zástupců Národní knihovny a dalších knihoven ČR. Nově je v samotném Formátovém výboru 15 trvalých členů (opět zástupci Národní knihov-

6 https://visk.nkp.cz/

7 https://www.ndk.cz/formatovy-vybor-ndk

ny a dalších knihoven) a na něj jsou navázány tři pracovní skupiny pro jednotlivé typy dokumentů. (6) Národní knihovna tak při zajišťování standardizace pro dlouhodobou archivaci spolupracuje s dalšími institucemi a snaží se využít jejich odborné znalosti.

Přestože na oblast digitální archivace je již v současnosti myšleno v rámci strategických a koncepčních materiálů, praktická ochrana dat, která v knihovnách vznikají je zatím stále na počátku. Stále více institucí si uvědomuje, že přenechat péči o svá data na jiných institucích není cesta, kterou by se chtěly vydat, a v rámci českého knihovnictví se objevuje více knihovníků zaměřených na tuto problematiku. V knihovnách také neustále narůstá objem digitálních dat, které spravují, stejně jako nároky na zpracovávání nových typů dat.

Použité zdroje:

1. *Státní kulturní politika na léta 2015 – 2020 (s výhledem do roku 2025)* [online]. Praha: Ministerstvo kultury, 2015 [cit. 2018-09-28]. Dostupné z: https://www.mkcr.cz/statni-kulturni-politika-69.html
2. Strategie Evropa 2020 – Digitalizace kulturního obsahu. *Ministerstvo kultury České republiky* [online]. Praha: Ministerstvo kultury [cit. 2018-09-28]. Dostupné z: https://www.mkcr.cz/strategie-evropa-2020-digitalizace-kulturniho-obsahu-831.html
3. Koncepce rozvoje knihoven ČR na léta 2017 – 2020. *Ústřední knihovnická rada ČR* [online]. Praha: Ústřední knihovnická rada, 2012 [cit. 2018-09-28]. Dostupné z: http://ukr.knihovna.cz/koncepce-rozvoje-knihoven-cr-na-leta-2017-2020/
4. Koncepce rozvoje knihoven ČR na léta 2011 – 2015. *Ústřední knihovnická rada ČR* [online]. Praha: Ústřední knihovnická rada, 2012 [cit. 2018-09-28]. Dostupné z: http://ukr.knihovna.cz/koncepce-rozvoje-knihoven-cr-na-leta-2011-2015-/
5. *Standardy digitalizace* [online]. Praha: Národní knihovna ČR, 2015 [cit. 2018-09-28]. Dostupné z: https://www.ndk.cz/standardy-digitalizace
6. *Formátový výbor NDK* [online]. Praha: Národní knihovna ČR, 2015 [cit. 2018-09-28]. Dostupné z: https://www.ndk.cz/formatovy-vybor-ndk

Long-term digital preservation in Poland: CREDO designers experience

Piotr Palka, Tomasz Traczyk, Warsaw University of Technology, Warsaw, Poland

Abstract

Long-term digital preservation is the process of maintaining digital objects through time, ensuring continued access to the objects. The CREDO project, the framework for Digital Documents Repository, is a digital repository which enables short- and long-term archiving of large volumes of digital resources. It acts both as a secure distributed file storage and as a digital archive, that provides metadata management and packaging resources in archival packages. Reliability of information readouts is ensured by the repository through the data replication and monitoring mechanisms in the repository file system, as well as through the distributed nature of the system that enables storing copies of the resources in more than one location. Advanced management system sup- ports scheduling of operations on the archival storage while respecting the low energy consumption requirements.

As a part of the CREDO project we also analyzed various methods of distributing a digital archive, from a simple dislocation of replicas, through solutions based on a cloud or blockchain technology, to the co-operation of federated archives and the use of multi-agent systems. Our analyzes are carried out in relation to requirements set for long-term archives by the OAIS reference model, resources security, potential costs for the user, real possibility of implementation, performance, and real solution volume. We would like to share the experience we gained while working on the CREDO project, in particular those related to the design of the repository, its launching and testing. We also got to know the point of view of potential customers.

Keywords: digital preservation, digital archiving, long-term preserva- tion, repositories management, OAIS

1 Introduction

Long-term digital preservation, the process of maintaining digital objects through time to ensure continued access, has become a crucial issue in recent years. The amount and the areas of digitized information are constantly increasing result- ing in obsolescence of the software and hardware required to preserve digital information. Despite recognized need for preservation action, still more work is required to effectively address the issue in theory and practice. Long-term preser- vation of digital information differs in many aspects from well-known short-term data storage. The longevity of the preservation causes a lot of little-known and difficult problems, which cannot be completely solved by technologies currently available, and some cannot be solved at all using only technical means. The prob- lem is of particular importance because of increasing prevalence of 'born digital' objects, which have no other representation than digital, so they are seriously threatened with the total loss.

In the paper one can find: (i) the state of affairs in Poland in a perspective of long-term digital data storage, (ii) the analysis of bibliography on distributed long-term repositories; (iii) the description of the architecture of CREDO digital repository, designed for trustworthy storage of large volumes of digital informa- tion; and (iv) the experience of CREDO archive designers.

2 Needs for long-term digital archiving in Poland

Poland, as many other countries, needs a number of different propositions for long-term preservation of different kinds of data, both analog or digital born. To mention only few: medical documentation, financial and accounting documen- tation, artwork, books, audiovisual data, web pages, etc. Each of this specific data types needs different assumptions for preservation. In Poland there exist both institutions and entrepreneurs interested in data archiving, and projects for long-term preservation.

2.1 National Digital Archives

The National Digital Archives (NAC), established in 2008, are one of three cen- tral archives of the state archive network in Poland. NAC is a response to the advancement of the technology of recording and providing access to data, estab- lished as a result of

the transformation of the Archives of Audiovisual Records. The operations have two aspects: digitalization and electronic documentation as well as storage of and providing access to photographic and audiovisual docu- mentation [2].

NAC digitize materials from the state archives across Poland. They build IT systems and infrastructure aimed to collect and provide access to the information on the collections of all state archives and other institutions storing archive materials. NAC collects, stores, maintains and processes photographs, films and sound recordings. NAC provides the audiovisual and digital archives, with 15 million photographs dated from the 1840s to present, and about 2,400 film titles from the period 1928-1993 [2].

2.2 Electronic Document Repository

National Library of Poland [7] in 2009 establishes Electronic Document Repos- itory, which collects publications disseminated by publishers in Poland only in the form of a digital file. To the publisher's repository, fulfilling the obligation stipulated in the Act on obligatory library copies, they provide documents saved in PDF format, such as e-books, electronic magazines and sound recordings. In addition, publications issued on physical media are transferred to the repository for secure archiving. Due to the fact that the objects collected in the Repository are mostly covered by copyright law, they are made available only at terminals in the National Library building.

2.3 Polish Security Printing Works (PWPW SA)

In December 2013, PWPW SA [12] signed a contract with the National Center for Research and Development (NCBiR) for co-financing a research and develop- ment project entitled Digital Document Repository (CREDO). PWPW imple- ments it within the framework of a consortium established, which also includes Warsaw University of Technology and Skytechnology company. The aim of the project is to develop and launch a demo version of a repository of 2 PB docu- ments, enabling efficient and safe storage and archiving of large digital resources in both short- and long-term periods.

The CREDO project meets the challenges associated with long-term storage of very large data sets generated by key institutions responsible for administra- tion, health service or culture. The scale and importance of the issue is huge, because the scope of work involves designing a system for storing data for up to 50 years, guaranteeing that

the archive will not only get the correct file when it comes to its structure, but also its effective opening or reconstruction (in the case of a media file), giving the real possibility of its further use. The potential for this type of service in the near future is estimated at dozens of petabytes and this number will dynamically grow in the following decades.

CREDO repository is described in Section 4. More details one can find in earlier publications [17, 8, 9].

2.4 National Data Storage System

A National Data Storage System (in Polish Krajowy Magazyn Danych) was avail- able via the PIONIER scientific network and MAN city networks. The system ensured reliability and security of data storage as well as high efficiency of ac- cess to them. The system's services were to be characterized by high availability, thanks to geographic relocation, infrastructure redundancy and internal failure mechanisms.

The main assumptions regarding the project were to ensure system scalabil- ity in many dimensions: capacity, performance, number of users served, number of system nodes and functionality, including the possibility of providing vari- ous methods of access to the system. Meeting these assumptions required de- veloping a dispersed infrastructure without a central point of failure, potential performance bottlenecks, and ensuring data and metadata replication as well as redundancy of infrastructure elements. Extensibility of functionality required the definition of standard, open interfaces between layers and system modules.

The project aimed to provide a reliable, secure and permanent storage of backup and archive data as well as efficient access to data for the needs of the scientific community in Poland as well as state and local government institutions. Unfortunately, the system is no longer maintained, although it was operating in 2012.

3 Digital archives – state of the art

Though digital data archiving becomes one of the most challenging problems of contemporary information technology, there are only a few fairly complete solutions. Relatively many products are offered for a short-time storage of large amounts of data.

Several hardware and software solutions are available, mainly suited to backup big data volumes in dedicated (e.g. MooseFS, ceph) or cloud (e.g. Amazon S3, Google Cloud) storage. Although these solutions are capable to store enormous data volumes, they do not address long-term archiving problems mentioned above.

3.1 Archives based on dedicated hardware

The most promising solutions for long-term archiving, e.g. [3, 13], are all based on OAIS standard [1], which addresses most of "logical" problems of digital archiving. Most of existing long-term archiving solutions are based on tape stor- age. However, promising disk-based, energetically efficient solutions also exist, like our CREDO archive, or Archi-Clarin project [6] based on specialized hard- ware. There are also long-term cloud-based solutions, e.g. [13]. Many important problems of information archiving, especially on long-term horizon, cannot be solved without preserving and managing metadata. It is widely accepted, and confirmed by OAIS standard, that metadata should be stored packed together with the described information into so called archival packages [1, 10]. Though metadata management is a key issue in context of archiving, the problem itself is very complex [11], and it is beyond the scope of this paper.

3.2 Blockchain-based archives

The world is developing solutions based on blockchain technology, which can be identified as long-term archives: three pilot solutions for the storage of land and mortgage registers (Brazil, Sweden and Honduras) and one solution for storing patient's medical records (Estonia). These solutions have been loosely categorized in [5] for following solutions: Mirror Type, Digital Record Type and Tokenized Type.

In the Mirror Type system, the blockchain serves as a repository for the fingerprints of digital records. The hash functions of digitized resources are cal- culated, which can be resources created as born digital or analog ones. In this way, a kind of record of digital shortcuts is created. These fingerprints are an- chored in a blockchain string that is used to check the integrity of records. This technology has been implemented in the Brazilian land and mortgage registry system and in the Estonian system of storing medical records. These are solu- tions that mimic the fingerprints of existing data. Regarding the capacity of such an archive, the Estonian system creates ca. 40,000 documents a day, in addition, logs about 1 million document transactions (browsing,

adding, changing, etc.). The system has a size of 2 TB (as of May 2017) with an annual increase of 300-400 GB [14]. Storing shortcuts of digital records in the blockchain provides the opportunity to check whether these records have not been corrupted, but questions whether such archive meets the requirements of OAIS.

In a solution called Digital Record Type, records are not a reflection of the digital fingerprints in the blockchain, but are actively created in the form of intelligent "contracts". For example, in the case of the Swedish pilot land and mortgage registry system, shareholders in the network include: property buyers and sellers, mortgage bankers and the Swedish land registration office. Conclu- sion of the transaction results in updating the distributed database of land and mortgage registers. This solution uses the postchain architecture, which is a kind of hybrid, combining the advantages of traditional database and blockchain tech- nology. It allows storing, downloading and modifying data, while ensuring data dispersion and redundancy of stored records using blockchain technology [4].

In the case of the Tokenized Type solution, not only records but also assets are stored in the chain. These assets can represent anything that has any value – land, artworks, etc. As the traditional currency transformed into cryptocurrency, and the land, wine, artworks, diamonds and other material objects, though still physically existing, can be "tokenized" or dematerialized in the form of virtual tokens that represent their physical form. In this way, in a tokenized blockchain storage system, literally every "thing" potentially becomes a record. This so- lution has been implemented in the pilot land registry system in Brazil. This solution also enables the storage of so-called blockchain born data, which were originally created in a blockchain network, such as all cryptocurrencies.

Archiving using blockchain technology is very promising, especially in light of the implementations described above. Data security is ensured by dislocation of blocks in a blockchain network, security of access is ensured by data encryption. The problem will be the long-term nature of the solution: it may not be possible to automatically change the format that is becoming obsolete. The capacity of the archive will be several or even several dozen times larger than the capacity of stored resources, but thanks to the fact that blockchain records are stored in the peer-to-peer network, this defect is no longer relevant. A certain obstacle in the application of such solutions may be the lack of control over the number of machines storing records and their high volatility.

3.3 Cloud-based archive

Cloud storage is currently very popular. Providers of these solutions offer, be- sides the obvious storage of data and access from anywhere, encryption func- tionality. In addition, data is replicated in several (depending on the solution) copies, which ensures data security. In addition, some clouds use data storage in distributed locations, which improves reliability in the event of a local failure.

Work [16] describes the use of the Microsoft Azure cloud for long-term storage of medical data (images). The author predicted that in 2014 in the USA, 100 PB of data from 1 million medical examinations was created. He proposes to use the Microsoft Azure cloud for archiving medical data. The use of a specialized cloud for archiving is promising, however, one should consider the energy efficiency of such a solution (traditionally clouds are available online) and migration and format diversification (these processes should be managed). In our opinion, it would be feasible and not very difficult to implement CREDO repository using cloud media instead of dedicated hardware.

3.4 Multi-agent based archives

The multi-agent system [18], [15] consists of many independent, communicating software units (agents), each of which has its own goals and information. It seems to be possible to define an archive as an agent whose purpose is to ensure the security of stored resources, and the information they possess is precisely those resources. In this case, a multi-agent system can be treated as: (i) a federation of closely related archives or (ii) a federation of loosely bound archives. Agents communicate to: exchange the resources for replication, exchange information to secure resources in the event of dangerous operations and exchange resources to recover a lost resource. Such functionality was considered in the CREDO project, and the proposal to implement such a solution was described in [17].

On the other hand, the agent may represent a certain set of archival re- sources. Then its goal is to ensure a reliable archiving of this collection. In this case, the multi-agent system is a distributed archive, as a collection of agents storing archival resources. Such a system contains a lot of agents, each of whom manages its own resources, dislocating them in the available infrastructure. Pro- viding appropriate information exchange protocols, individual agents can store resources in any way they want (e.g. in the cloud, using blockchains, in a classic store, etc.). The management of migration

and the diversification of resources would also lies on each agent. Due to the autonomy of agents, one can imagine that each of them has any convenient configuration. When the archival pack- age is to be disseminated, you would have to find the agent responsible for the specific package.

4 CREDO overview

The CREDO[1] system is a digital repository which enables short- and long- term archiving of large volumes of digital resources. The project was led by Polish Security Printing Works [12], in cooperation with Warsaw University of Technology.

By design the repository is to act both as a secure file storage and as a long- term digital archive, providing metadata management, and storing digital re- sources packed together with their metadata in archival packages. When acting as an archive, repository meets requirements of the OAIS (Open Archival In- formation System [1], ISO 14721:2012) standard that provides recommended practices for implementation of digital archives. One of the primary functions of the repository is the support for various currently available data carriers: hard drives, solid state drives, tapes. Reliability of information readouts is ensured by the data recording replication mechanisms in the used file system, as well as the distributed nature of the system that enables storing copies of the resources in remote locations. The repository architecture is multi-tiered and consists of loosely-coupled subsystems, which enables, together with the emergence of new technologies, replacement and continuous upgrades of individual components.

4.1 Archiving issues

The CREDO archive addresses most of the main long-term archiving problems.

– A technology obsolescence is addressed by possible utilization of various filesystems, based on various media, and by relatively easy replacement of system modules.
– Information readability is ensured by storing digital resources only in selected, sustained and non-proprietary formats, recognized as well-suited for digital archiv-

1 CREDO – the acronym of Polish name Cyfrowe REpozytorium DOkumentow, which means ‘Digital Document Repository’.

ing, together with technical metadata describing the resources, and documentation of formats used.
– Information intelligibility is partially facilitated by storing descriptive metadata of the archived resources.
– A storage uncertainty is addressed by:
  • regular multi-level monitoring of data correctness, with use of low-level (built in filesystem) and high-level CRC checks;
  • multi-level data replication: low-level replicas automatically maintained by filesystem, and high-level replicas maintained by CREDO archive;
  • automatic data dislocation between different locations,
  • limited support for dislocation into separate federated archives, includ- ing synchronization of activities on resource copies stored in separate archives;
  • continuously diagnosing and monitoring of data carriers and media replacement management.
– Economic efficiency is ensured by scheduling algorithms which appoint the schedule for turning the servers on, when all the data storing servers are normally in the off state.

4.2 CREDO architecture

The CREDO Digital Document Repository is divided into several subsystems (see Fig. 1).

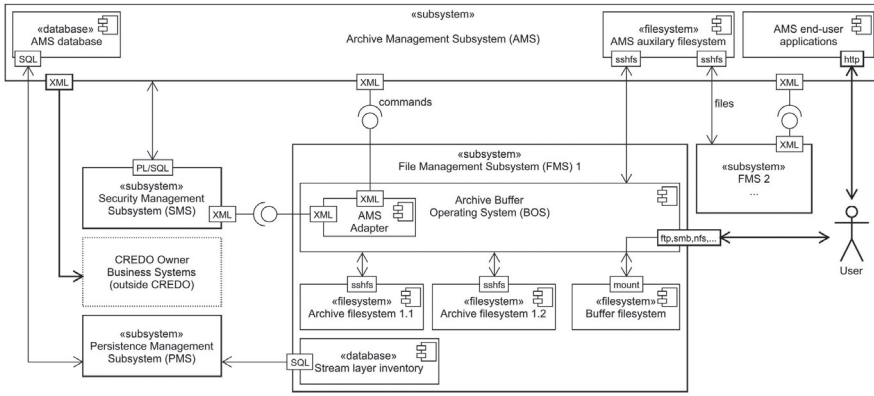


Fig. 1. CREDO Digital Document Repository architecture

– Archive Management Subsystem (AMS) controls activities of the archive: ingest, search, and outgest sessions (as defined in OAIS standard [1]), inter- nal administrative processes, etc, and interacts with end-user. The subsystem stores also in its database partial metadata of archived resources. This meta- data enables fast search in the archive and is used to verify archive integrity and authenticity.
– Persistence Management Subsystem (PMS) monitors archive storage and works out recommendations for data allocation and relocation, ensuring low storage failure risk. It also optimizes the operation of the storage media to achieve energy savings. It co-operates with AMS, providing schedules of AMS activities.
– Security Management Subsystem (SMS) controls access to archives filesys- tems and the archive buffer. It interacts with AMS to grant access rights necessary for AMS activities but, for security reasons, it is separated from AMS. If necessary, it can be implemented in different technology and run on physically separate hardware.
– File Management Subsystems (FMS) consists of buffer filesystem, archive storage filesystem and some auxiliary modules. One repository can use many such subsystems, even implemented differently, but interacting with other subsystems of the repository with the same set of well-defined protocols. Each FMS can contain many filesystems, which may also be technologically diversified.

Due to the planned longevity, the repository cannot be dependent on any par- ticular technology, as hardware and software. This is ensured by layer-component system architecture with loose coupling between subsystems, and clearly de- fined interfaces between them, which use standards-based communication protocols: RESTful Web Services containing self-describing XML messages, and SQL queries (which also may be replaced by REST/XML if necessary).

4.3 Persistence management subsystem

Persistence Management Subsystem (PMS) sets guidelines for the stream layer (file system) on replicas placement, relocation of data, data carriers diagnose and replacement, scheduling the access to the archive, and power management.

Replication Replication is the data migration method, that does not change the bit sequence on the data carrier. It assumes creating multiple copies of the same resource. Additionally, the replication assumes that the resource is copied onto another carriers. In CREDO, there are two levels of replication: low-level replication, that is assured by the file system; and high-level replication, provided with the PMS. Both replication

methods serve to ensure data security. Moreover, the high-level replication can be assured on: (i) area level, where different copies of the same resource are kept under different paths, (ii) file system level, where a resource is stored on different file systems, (iii) archive level, where a resource is secured on a few archives.

The higher the level of the replication, the data is safer, as the mechanisms ensure: doubling the resource safety, diversification of a resource representation (when the copies are kept under different file systems), geographical dispersion, when resource is stored in different locations.

Relocation The goal of the relocation is to prevent data aging by refreshing the resource on another: carrier, area, file system, or archive. PMS assumes package relocation using optimization methods, where the packages are relocated: peri- odically, on safer area, with sub-optimal areas usage, and with energy efficiency. Moreover, the relocation is coupled with the carriers replacement. The packages are automatically high-level replicated among the areas, the PMS schedules the plan for a long-period.

Data carriers diagnose and replacement The goals of the diagnose and replacement of data carriers module are:

– Analysis of a risk concerning failures of data carriers and whole areas de- signed for storage of archival packages. The analysis consider single pieces, lots, models of carriers, and it goal is to predict the failure, and secure the data stored on the carrier by replication, relocation on the other carriers, renewing, or emptying it.
– Replacement of carriers decision support system.
– Failure prevention due to prediction of the failure.
– Interoperability with the data relocation, by automatic rewriting data from the carriers under threat.

To obtain the goals, the module for carriers, and area reliability assessment is used. It is common mechanism to provide the information about different: carriers, areas, batches of carries to PMS. Moreover it provides information in- dependently of carrier type (hard disk, magnetic tape, CD, DVD, Blue-Ray, pen-drive, etc.). Finally it solves the problem of a lack of technical solutions for monitoring, and managing carriers. The module allows to choose a source or destination for the archival packages during archive operation. Also, it specifies the moment for the migration. Finally, it determines the areas for powering-on during reading or writing the archival packages.

Power management In CREDO the carriers are divided onto storage ar- eas, that are subject to the PMS management. Single storage area has assigned a number of data carriers. The allocation of the packages onto storage areas is done by PMS, and a power management module manages areas starting-up, and shutting-down. Power management system is coupled with a scheduling module, and it have to start-up given area, when the schedule assumes operating on it. Similarly, when planned operations finish, the power management module shuts the area down.

5 Designing long-term digital archive – CREDO designers experience

Our experience, after completing the task of designing and implementing a work- ing long-term archive, is invaluable. In such an archive, it is important not only to securely store large data volumes along with complex metadata, but also to ensure long-term persistence of data packets, energy efficiency, and to provide the reliable mechanisms for periodic migrations to more reliable carriers or to new formats, checking of a lot of package parameters. Therefore the task seems to be an interdisciplinary problem, which needs appropriate skills in fields of IT, archiving and librarianship, optimization, and even physics and chemistry.

As the problem of digital archiving is relatively new and almost none proven solutions are widely accepted, it was quite difficult to set the right assumptions of the project. In many cases the originally accepted assumptions had to be changed due to the accumulated experience. Sometimes seemingly small changes, such as the decision to apply an additional buffer between the archive and the client, resulted in significant changes in the operation of the archive and the level of data security.

One of the most difficult and important problems related to digital archives is ensuring energy efficiency. As the CREDO system is mainly based on hard disk storage, this problem was particularly important. A module was designed for scheduling access to media, ensuring that the memory power supply is switched on as rarely as possible. It required use of complex algorithms in the field of operational research, and a specially designed state machine, controlling archive operation in cooperation with scheduling algorithms.

Ensuring long-term persistence of digital data requires control over the po- tential unreliability of the data carriers, to enable migration to more reliable carriers before failures occur. The development of mechanisms to predict media reliability required deep entry into the theory of reliability and into chemical ki- netics. Importantly, designed mechanisms are closely related to the specific type of media, because each media type requires different knowledge.

Metadata management is one of the most important issues in digital archives, as the stored resources can become useless if their metadata are lost. There are many complex problems related to metadata, resulting from multiplicity of used standards and formats, different expectations of various user groups (e.g. librar- ians versus content creators), various levels of accuracy of available metadata, labor consumption of metadata introduction, etc. The solutions adopted had to be flexible enough to provide support for a variety of formats while maintaining the relative ease of entering and transforming information. The use of XML- based formats along with universal tools for processing this type of data (XSLT, XQuery) proved to be a good choice.

One of the most difficult issues solved in the CREDO project turned out to be the processing of information in accordance with the requirements of the OAIS standard [1]. Although this had not been foreseen in the original project assump- tions, the business partner demanded that such functionality be introduced into the repository. This required the creation of appropriate mechanisms for man- aging the processing of archival packages, which interact with the optimization of energy consumption.

Although the majority of technical problems related to the operation of the long-term archive have been solved, the implementation of the created system for practical use is still far away. However, problems with implementation result mainly not from substantive/technical issues, but from difficulties in cooperation between the university and business. These difficulties result i.a. from a lack of understanding of the role that universities should play in research and develop- ment projects. Business partners expect scientists to create ready-made IT solu- tions, while the university is not a software-house: its role is to solve significant problems, not programming. Another source of problems in the implementation of such complex system is the instability of the business partner strategy, result- ing i.a. from personnel changes at the business partner, and even for political reasons. Therefore, it seems that the system designed directly for the final user, such as the state archive, which has stable goals and needs, would have a better chance of effective implementation. In the case of the CREDO project, the busi- ness partner was an intermediary who intended to sell the ready-to-use sys-

tem to end users. Both the ambiguity of the business goals of such an intermediary, and the difficulty in acquiring long-term customers, significantly hamper the success of the project implementation.

6 Summary

Storing the data for dozens of years introduces many issues, to mention only the most important: technology obsolescence, energetic efficiency, file format expi- ration, hardware and operating systems out-dating. Poland, as other countries, needs the solution for archiving different kinds of data: both analog and digi- tal born data, as audio-visual data, medical information, documents, web pages, etc. The literature offers us some insights for developing the long-term digital document repository; in this paper and our other publications we described our CREDO archive.

CREDO Repository is designed to trustworthy and cost-effectively store large amounts of digital resources (the instance built has over 2PB capacity), with use of standard hardware and data carriers available now, and in the future. The so- lution addresses most of main problems of long-term digital archiving of digital resources, including possibility to adopt new technologies, and flexible metadata management, which should ensure information authenticity, readability and in- telligibility.

The solution is designed for institutions that store large digital resources for long periods of time, e.g. institutions responsible for cultural heritage, mass media, state administration offices, health care institutions, etc.

Acknowledgments

The project entitled Cyfrowe Repozytorium Dokument´ow CREDO (Digital Document Repository CREDO) is co-financed by the European Union through the European Regional Development Fund under the Operational Programme ’In- novative Economy’ for the years 2007–2013, Priority Axis 1 – Research and development of modern technologies, Grant No. WND-DEM-1-385/00.

References

1. Consultative Committee for Space Data Systems. Reference model for an open archival information system (OAIS). Recommended practice., June 2012. Access: 2016-10-25.
2. National Digital Archives (Narodowe Archiwum Cyfrowe). http://www.nac.gov.pl/. Access: 2018-09-18.
3. David Giaretta. Advanced digital preservation. Springer Science & Business Media, 2011. ISBN 978-3-642-16808-6.
4. Victoria L Lemieux. Evaluating the use of blockchain in land transactions: An archival science perspective. European Property Law Journal, 6(3):392–440, 2017.
5. Victoria L Lemieux. A typology of blockchain recordkeeping solutions and some reflections on their implications for the future of archival preservation. In Big Data (Big Data), 2017 IEEE International Conference on, pages 2271–2278. IEEE, 2017.
6. K. Marasek and J.P. Walczak. Long-term preservation of digital files in data network structures (in Polish). http://www.ci.pw.edu.pl/content/download/1426/11818/file/KMJPW06102015-fin.pdf. Access: 2015-12-01.
7. National Library of Poland (Biblioteka Narodowa). https://www.bn.org.pl. Access: 2018-09-18.
8. Piotr Palka, Tomasz S´liwin´ski, Tomasz Traczyk, and W lodzimierz Ogryczak. Per- sistence management in digital document repository. In Beyond Databases, Archi- tectures and Structures. Advanced Technologies for Data Mining and Knowledge Discovery, pages 668–682. Springer, 2015.
9. Piotr Pa lka and Tomasz Traczyk. Metadata processing in CREDO long-term digital archive. Studia Informatica, 38(2):79–91, 2017.
10. K. Pater and T. Traczyk. Opakowanie zasobow cyfrowych na potrzeby archiwizacji dlugoterminowej. Studia Informatica, 34(No 2B (112)):898–103, 2013.
11. G. P loszajski (ed.). Standardy techniczne obiekt´ow cyfrowych przy digitalizacji dziedzictwa kulturowego. Biblioteka G lowna Politechniki Warszawskiej, Warszawa, 2008.
12. Polish Security Printing Works (Polska Wytw´ornia Papierow Warto´sciowych). http://www.pwpw.pl/. Access: 2018-09-18.
13. S. Rabinovici-Cohen, J. Marberg, K. Nagin, and D. Pease. PDS Cloud: Long term digital preservation in the cloud. In Cloud Engineering (IC2E), 2013 IEEE International Conference on, pages 38–45, March 2013.
14. Records in the chain project, estonian e-health system (rcpeu-01) case study 1, 2017.

15. Y. Shoham and K. Leyton-Brown. Multiagent Systems: Algorithmic, Game- Theoretic, and Logical Foundations. Cambridge University Press, 2008.
16. Chia-Chi Teng, Jonathan Mitchell, Christopher Walker, Alex Swan, Cesar Davila, David Howard, and Travis Needham. A medical image archive solution in the cloud. In Software Engineering and Service Sciences (ICSESS), 2010 IEEE International Conference on, pages 431–434. IEEE, 2010.
17. T. Traczyk, W. Ogryczak, P. Palka, and T. S´liwin´ski. Digital Preservation: Putting It to Work, volume 700 of Studies in Computational Intelligence. Springer International Publishing, 2017.
18. Michael Wooldridge. An introduction to multiagent systems. John Wiley & Sons, 2009.

From the Past to Eternity – Long Term Preservation Derived From a Collaborative Platform such as FOLIO and New Data Models such as FRBR

Miklós Lendvay, National Széchényi Library, Budapest, Hungary

Abstract

The past provided a lot of materials for libraries to preserve. But today, the amount of publications created in all fields of culture is rapidly growing to a never seen extent. What is even a bigger challenge, the amount of machine-generated and unstructured data is growing at a rate more than exponential. If we want to face up to the challenge of preserving some of these data for long term, without drowning in the sea of data, we need to create new ways of cooperation when selecting, structuring and processing data. This inevitably leads to a completely new data exchange and cataloguing model, a cooperation platform and framework enabling more flexible workflows in data processing and opening up the possibility of cooperation for a much wider audience: citizens, scientists, and institutions.

This is why the National Széchényi Library took a leading role in defining a cooperative and distributed nationwide library platform and is developing a self-hosted cloud-based environment. Our library is committed to participating in the FOLIO (The Future of Libraries is Open) community and contributing to developing the software with the results achieved when creating the platform.

How does the new data philosophy FRBR (Functional Requirements for Bibliographic Records) and the new data exchange format BIBFRAME (Bibliographic Framework) affect the process of archiving data? To what extent can the FOLIO platform enhance the creation of a meaningful, new archive technology? We are yet to see the answers to these questions.

The challenges libraries have to face today are manifold. The biggest challenge resulted from the extensively growing number of publications and the amount of information that the library has to collect, process and make available.

The possibility of publication is not a privilege for the chosen few any more, as it used to be in the past, but is available for almost all the people. These days, only a fraction of the published materials goes through traditional channels such as editorial systems, publishing houses and alike. Private publishing is no longer a rarity. However, selecting received materials, carried out in the past by well-organized editorial systems and peer reviews, is an overwhelming task and this raises a question: Are national libraries to take on a new role, and take responsibility for selecting materials and judging the content of these? Prioritization, selection and filtering of the received materials is an absolutely must – are we allowed to do that? And if we are, to what extent we should do that? The intention is to collect and archive everything, and this is also anchored in the prevailing law. Where are the limitations of collections? Who in the library should select the scope of incoming materials?

The second biggest challenge today is the variety of materials. Earlier, printed materials dominated, but today the amount of online publications and the number of digitally created materials is increasing, in an abundance never seen before. Also, new types of materials, such as video recordings of scientific lectures, web archives etc. are arising, changing the library landscape.

A further challenge is that users of the library, patrons are accustomed to the possibility of actively contributing and not only consuming in a widely shared environment. These users demand that they have similar opportunities also in the library sector – which, if it ever happens, once again would increase the number of published materials.

In existing library systems we have a huge amount of data, and in quite diverse formats. The legacy is growing by attaching new parts and new fields to existing structures. The dominant data format in libraries is MARC21, originating from 50 years ago when the intention was to have a machine-readable data exchange format (hence the name: Machine Readable Code or ‘MARC’). Since then, other national formats have been derived from MARC21 (USMARC, HUNMARC etc.), but the overall aim is to have a unified data structure that makes sense for all participants. For unknown reasons, in most cases software systems themselves have limited the internal data format to exchange formats, resulting in much poorer internal descriptions than necessary. In the last decades as new demands for fields emerged, information was added to the

linear structure of the record, making it bigger and bigger: now we are very close in number to two thousand fields and subfields in a single MARC record.

The library sector has created a big variety of software products and systems. These are mostly software systems, which, similarly to MARC are constantly growing, responding to new demands as elements are added to existing structures. The result is big monolithic systems, where flexible changes or additional features are very hard to incorporate. Often the cost of the change cannot be justified, and the user libraries in their everyday work have to take the limitations of such golems into account, without hope of realizing their demands. Small library systems vanished from the market, the tendency is that only a few monolithic systems survive in the competition. For many libraries, it is financially not feasible to purchase a new system and are developing in house. They do this with the help of a very few (mostly underpaid) software developers, using internal systems and, in many cases, open source software solutions. The result is a patchwork of poorly integrated (sometimes not at all integrated) software pieces which have no clearly defined standard interfaces to interact with each other, and thus, cannot use central functions such as user authentication, access management, workflow engines etc.

Traditional systems and data models influence and, in a certain respect, determine – and freeze – workflows in libraries. Our rigidly structured, old-fashioned hierarchies, are formed by the good old punch-card logic and use traditional catalogue cards for printed materials, mostly for monographic books. This library structure mirrors the way libraries have been functioning in the last centuries. There is little or no scope for new ways of collaborations technology would otherwise enable; the walls of these institutions are like medieval forts, and scientists and experts not integrated into an institution are cut off from the possibility of contributing to the library catalogues and materials.

In the meantime, IT systems in our everyday life are characterized by many new features: flexibility, individuality and commonly shared zones, connectedness, linked data, structures and workflows for sharing the data, microservices, collaboration and so on. Many obstacles and odd circumstances separate the new generation users from the treasures of libraries. It is very rare that libraries are able to offer users 21st century electronic services patrons outside of the library domain are used to. The world of the library is separated from the rest of the world, there is no real fluid exchange between the two. Users find access to library materials too complicated while the effort is not justified for them, in comparison to, for example, using Google search or other internet-based services.

New generations brought about a fundamental shift in society. Today`s youth have a mindset different from that of previous generations. Instead of old structures and rigid hierarchies of command and obedience, young people are used to existing and cooperating in networks. In these networks, participants have equal rights, the responsibility to contribute to a certain field is solely determined by expertise, and based on this, one can assume a leading role in a work context. Instead of being controlled by the hierarchical institutional positions, they draw on the wisdom of the community. They empower both each other and the working group as a whole by extended cooperation. They prefer taking small steps instead of planning far ahead into the future. They experiment, and the next reasonable step is guided by the result of such free experimentation, and processes are defined via community consensus rather than an appointed individual. They value openness, sharing, and transparency above all; dark chambers of secrecy are not appealing for them anymore.

In the last couple of centuries, knowledge and secrecy of information meant power for the privileged few, there is a new tendency for openness, enforced by young people. The young generation has the right and the obligation to redefine the existing society according their rules. It is also a remarkable tendency in religious communities: hidden secret sects and churches are today replaced by the openness of religious communities where even the role of the priest could be taken over by a member of community to announce information or act as a temporary leader. Secret knowledge thus becomes public knowledge.

What is known must be shared! This is also a motto of OCLC, the biggest library community, and this is also increasingly true in the context of library services. On a broader scale, demand for a certain level of freedom is present in the younger generation: they cannot be motivated by material wealth and profit, rather, they actively seek the fulfilment of their aspirations and above all they wish to see purpose and meaning in what they do. This shift of mindset is illustrated by Aaron Sachs's and Anupam Kundu's picture (Figure 1) – although making reference to the context of organizational change, the same applies to libraries as organizations and also to society in general.

Summarizing the characteristics of a library of today in three selected aspects: data, software solutions and work sharing, my investigation shows the following:

Data formats across systems are limited to exchange formats such as MARC or Dublin Core. These are linear descriptions, which means information is repeatedly described and the resulting entries are not connected to centrally described entities. Thus, with very few exceptions, linking of existing data is not realized. This results in data dupli-

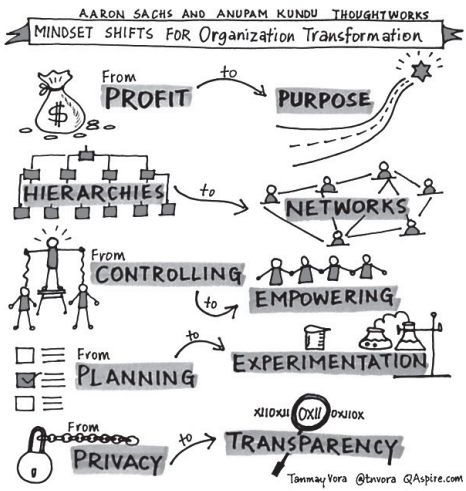


Figure 1.

cation as well as a duplication of effort. Any piece of connected information is not centrally connected to the main entry, but rather to any random catalogue entry, and this causes lost connections in the data cloud. Copy cataloguing is a common way of creating duplicate information – instead of using and linking existing information – where already existing information is duplicated and altered for the own purposes of the internal system. This data is mostly isolated in institutional systems, it is not public nor widely available for cooperators – except sometimes in a limited way for other libraries of a given community / consortia.

Library systems are monolithic solutions – one ever-growing big system which is constantly striving to fulfil new demands while using an existing logic and structure, mostly too narrow to satisfy expectations. Communication with the outside word is solved by the usage of a number of API-s, which again adds to the complexity of the system. The software used is mostly licensed, closed source, owned by development companies or the library communities.

Sharing of work happens mostly within an institutional hierarchy, with very little cooperation between institutions. Protection and privacy of materials is very important, copyright restrictions and the cautious European copyright approach (nothing is allowed, unless a written formal approval from the rightsholder is provided) makes a lot of materials only available inside the library building. Remote access, safe structures and complex copyright management are a rarity.

If we stick to this approach, growth in many areas will remain a curse for us. If we are not able to shift our mindset and to take a big step in organizing our work in a completely different way, we will drown in the sea of data. We cannot do justice to arising new sources and new players wishing to participate in this field, and definitely cannot handle new types of media such as photos, private electronic collections and web materials. Do we have this as a real option? Do we have a free choice? Can we afford to stay in our well known but limiting structures, or do we have to inevitably broaden our understanding and drastically change our thinking? Human beings are capable of this, any moment starting anew, reforming their attitude, all aspects of their behavior using their creative capabilities. Out of human always derives a choice.

What feasible approaches can be identified, as we fundamentally change our mindset and pay attention to the shift already happening in society? How would these three fields, data, software systems and work sharing / cooperation look like in this new paradigm? What is the renewed vision, towards which the National Széchényi Library and its collaborating partners (public, academic, church libraries) will strive towards in the near future?

The data model must be flexible, and completely the choice of any participating system. There should be no limitation whatsoever, how the data are described and in what ways the entities are connected. Only this approach can ensure, that the exchange and the maintenance of the data is not limited to the library domain, and can be expanded to any existing domain with its free and individual data formats. The link to cultural institutions like museums and archives is obvious, but there must be a possibility of cooperation opened with geographic name space institutions, with the domains of military, education, transport, production, design etc. The data sources should not be limited to institutions, as many experts in certain domains, many scientists are not embedded into institutional hierarchies. They should be capable of enriching the platform data, the doors of ingesting digital data and catalogue information must be open for the widest possible range of participants. The legacy of writers, with a lots of electronic manuscripts and not published materials, have to be able to find their ways into safeguarded and archived library platforms. Citizen science and crowdsourcing must be enabled. The data model should allow for any type of data format, must be definable in a flexible way, without programmers' intervention.

We have to aim to describe the entities by those experts, who have the deepest and most reliable knowledge about them. Such entities could be any, the mostly used in a library context are persons, institutions (libraries, publishing houses), geographic name spaces and chronological dates, the most important ones being works, instances,

agents. Those entities should be described only once and should be linked by connections to other entities. The enrichment should happen strictly by adding information with a very precise information about the source, above all reflecting the trustworthiness of the source. This approach will give to each entity and the relevant connections a quality level qualification, and this is a clear indication of trustworthiness. It allows to every participant to attach information freely and with an immediate effect, but for consumers and users there will be a choice of selecting the range of trustworthiness, as they work and utilize the selected data. In our project we will additionally introduce a module called Loca Credibilia (Trusted Source / Place), which identifies the source and originator of a digital object, allowing to trace back the route of a given digital object. All entries can have many versions, can have competing data entries, but even at individual field level the information of the source, and with this the trustworthiness, is provided and stored.

With the growing amount of data produced and waiting to be processed, it would be a naïve approach to assume, that the problem could be solved by the increase of the number of staff, hiring new cataloguers. The involvement of a broader expert community is a must, but we should not ignore the option of the machine supported processing. The machine can do processes on its own, without manual human intervention, and there are tools and means, by which participants can teach the machine workflow and increase significantly the accuracy of the computer's process. Many processes in improving the quality of digital materials are already making the life of colleagues easier, and this feature can be easily introduced also in cataloguing the web harvested materials, or in recognition of faces, objects etc. at the identification and cataloguing of still and moving images. The machine can give a huge support in improving data quality.

As the data is completely flexible and broken down to individual fields / entities, into the smallest possible units, the same way the software system elements must also be broken down to the smallest possible units, modules, and must be equally interconnected, allowing the choice of components for the system users. Instead of monolithic systems the modularity is a must! The above described data model needs distributed systems, where the participants are defined in a flexible authentication and authorization system, and flexibly defined control parameters will determine the workflows and the connections within the system for all participants. This approach allows a very unique complex setup for each co-worker, without programmers' intervention, by setting up the relevant parameters within the parameter setups. The machine/software needs to be able to retrieve the information about the ever changing context, where the user (librarian and patron alike) comes from, and according to this information the next

step can be derived automatically and the user can be supported in his work with the support of automated and fluent workflows.

There are many good examples of collaborative platforms, where the materials can be reached widely in a very individual way. My personal favorite is Spotify, which provides music for the interested audience. This system comes very close to what a library has to offer to its users. The songs and music pieces are well organized, can be accessed in many different ways. The authors are well-defined as entities, information about them, history, recent concerts etc. are widely provided. The access to the music pieces is on individual level, anyone can set up according his interest the playlists, can even start a new customized radio station (which is generated half by human, half by machine), and by providing information about preferences, can teach the system, what materials should be suggested for him. But this alone would be a stand-alone, lonely system! A wide range of collaboration can happen. Playlists can be shared, used and maintained (built) by friends, family or any community. Anyone can be followed: artists, who provide materials, and listeners, who consume information. Inspirations can be received by browsing genre, friend's listening habits, suggestions can be made to each other. The system automatically generates "Discover Weekly" suggestion, which is a weekly mixtape of fresh music, and enables new discoveries. "Release Radar" will detect the new contributions of the favored artists. What a pleasure to collaborate in such a delicate matter, as the music we are listening to!

Amazing, how Spotify develops the software. They realized that if you want to reach a 100 % predictability, you will end up with 0 % innovation. So if you are really looking for innovative systems, you need to accept the necessity of dead-end approaches and failures. They even encourage themselves to celebrate failure! It has its own learning benefits, as we try out something, and learn by doing and from own experience, if this is feasible or not. Being responsible developers, they create a fail-friendly environment, where a failure does not cause a catastrophe in the everyday operation, instead of avoiding failures they concentrate on failure recovery. Instead of having huge release changes, they take the step-by-step approach, and they minimize the headache of a release change by having small changes. Some key ideas of this development are: limited blast radius via decoupled architecture and gradual rollout, experiment friendly culture, waste-repellent culture, healthy culture heals broken process, culture-focused roles. This is in every aspect a 21st century approach. (Figure 2 and 3)

There is a huge potential in sharing the software with the community. The tendency of the modern era is the usage of open source systems, where the community can support the development of certain functions. The smaller the modules are, the biggest is the

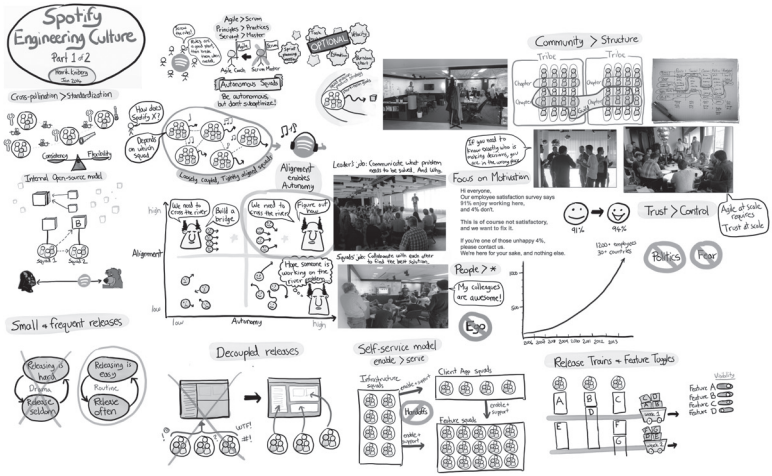


Figure 2.

chance that the modules can be interconnected in a very flexible and fruitful way. By setting the commonly accepted rules between the modules and the communication infrastructure, the freedom of choice is guaranteed: variations of the same module can be produced and made available, and existing modules can be customized if needed. The ideal solution is of course, if the processes are driven within flexible modules by control parameters and setups, without having separate modules for individual variations.

The sharing of work is a big chance and potential in overcoming the staff shortage (and often the shortage in domain specific expertise). This enhances the quality of data, and can result in a never seen productivity and richness of data. The model of cooperation must be open and transparent for every possible participant. The willingness of forming the word and the information systems is given, as young users are used to it, and this is time for the libraries to make use of this potential. An additional benefit is that the users who are contributing to the richness of the library, will very likely use the materials provided by the system and will come closer to the library domain. The separation of the open world and a closeness/secretiveness of the libraries can disappear, by integrating the library treasures into a broad collaborative system. The bridges to WikiData, WikiMedia, OpenStreetMap and alike can be built and used more extensively, and this will connect further users to the library domain.

If libraries can achieve this, the growth is not a curse any more, but a blessing. In the above context the growth contributes to the richness of well-organized materials. The hidden knowledge of experts can be brought to daylight, the machine supported pro-

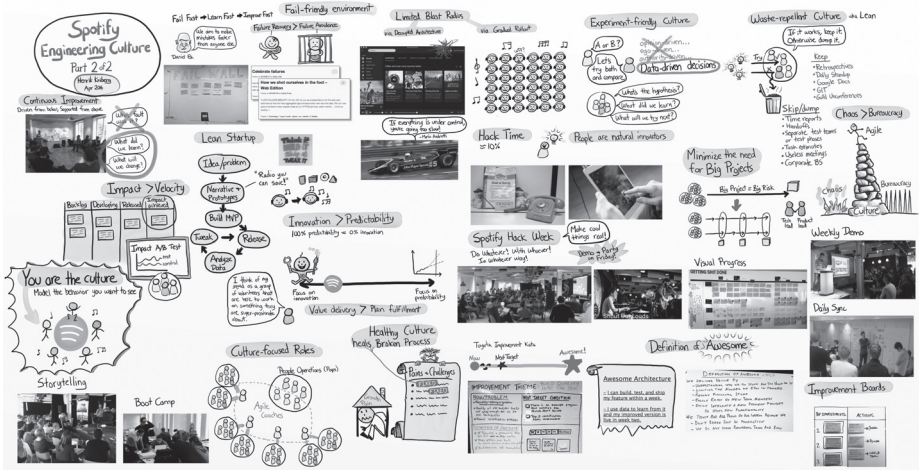


Figure 3.

cesses can result in an immediate accessibility of digitally born data, and the flood of data is beneficial for all participants. The freedom of choice is guaranteed in all aspects. The libraries can define their own responsibilities, workflows and such. The boundaries between institution types will be dismantled, the cooperation of institutions will not be limited any more. The users will be able to individualize their access regarding consumption, and regarding their contribution. The collaborative model will bear fruits easily and using the immediately available (mostly open source) technical solutions will harvest the low hanging fruits very soon.

There is a community developed platform, which pays attention to the above detailed aspects. The platform is called FOLIO, "The Future of Libraries is Open". This platform with its strict communication rules accepted and obeyed by the developers / contributors can ensure, that individually developed modules can communicate with each other. The technical heart of this is OKAPI, an API gateway that manages communication and separation between apps and different tenants (installations) of the platform. OKAPI is a message bus, the communication channel. The approach is based on microservices: the smallest possible units interconnected in a meaningful context. In this the applications are language agnostic, applications can be written in any programming language – which again provides the freedom of choice for the developer community. The apps can be open source, can be licensed, software developer companies can integrate their products. The cooperation in FOLIO is based on the work sharing of librarians, functionalists, strategists, service providers and developers, and enriched by active users.

The Hungarian National Library has signed a cooperation agreement with FOLIO in 2016. The first press release of FOLIO already announces a broad range of cooperation partners, among them our library. Since then, our library is an active member of the community, and with our requirements we aim to influence the scope of functionality and the openness of the designed system. We have organized early 2017 a FOLIO day in our library and invited all Hungarian libraries to learn more about this revolutionary approach. We have sent invitation to the neighboring countries as well, and have recorded the presentations, which are available on our website www.oszk.hu/en. Since then the community grew significantly, the World Of Open Libraries Conference (WolfCon) held in Durham (Northern Carolina) in 2018 has received already 180 active participants, who are actively engaged in the work with and around FOLIO.

In collaboration with other types of libraries (public, academic, university) the NSZL has worked extensively on the requirements of a new national library platform for Hungary. The result was a 70 pages document summarizing the broad overview of the system and about 1.100 lines of requirements in addition to the "normal" functions of already existing library systems like Alma, WordShare, OliSuit or Qulto. Our intention is to create a cloud based hardware environment where this new collaborative platform can run and be accessible for all participants. In the tender process it became very obvious, that existing library systems cannot fulfil the requirement the Hungarian library community has expressed, so the only feasible way forward is to implement the FOLIO framework, adding missing functionality to the existing modules and develop the modules missing in cooperation with the community members. (Figure 4)

As a preparation for the implementation we are cleaning and converting the existing data from our library system AMICUS and from all the existing systems around our ILS (we have many different and outdated solutions for digital objects, etc). In the meantime, with our external software developer partner we are in the process of developing the necessary software to handle entities for our new system (National Namespace), a flexible cataloguing module to be able to catalogue any (really any!) type of objects, a flexible access right module, and we are using and parametrizing an open source web harvesting system. These are all based upon the concept of our new FOLIO system using linked data, obeying the rules of the new cataloguing standard RDA (Resource Description and Access), the new library data exchange format BIBFRAME 2.0 (Bibliographic Framework) and the high level conceptual reference model developed within the entity-relationship framework LRM (Library Reference Model) by IFLA. All these data connections are based on the FRBR (Functional Requirements of Bibliographic Records) philosophy, which has its own but connected data model version in the museum and archive domains.

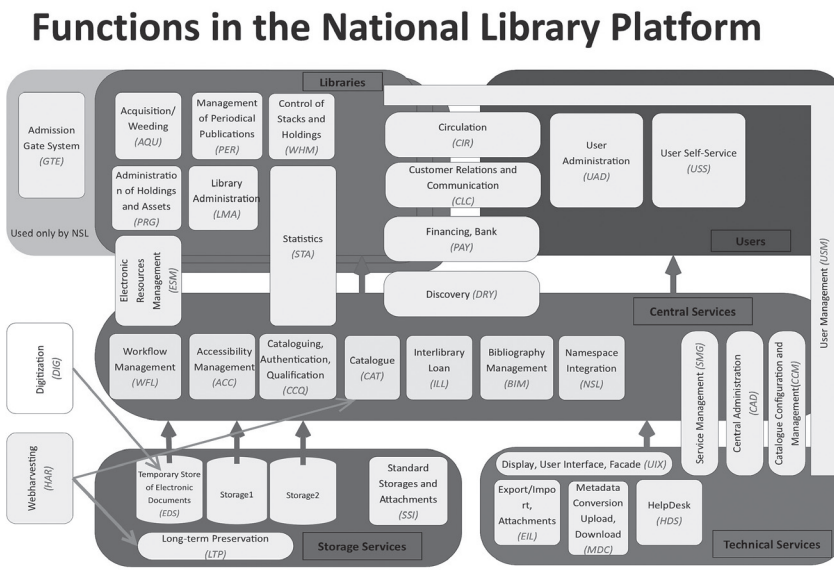


Figure 4.

The entities libraries use reach far beyond the boundaries of libraries themselves. Therefore, the namespace development in Hungary is a collaboration of many institutions, and although NSZL is the lead in the development and will provide the infrastructure for the system to run upon, there are further 11 institutions right now (and the number is growing) whoare considering to be linked and are ready to contribute to the system. The underlying principle is the language model: the complete Hungarian language with all Hungarian words in it is labelled with identifiers. This means, that any namespace created will use these identifiers, there are no strings applied in the system – and with this the system can be language agnostic, if other languages are connected to the identifiers. So far the following namespaces are created: person, corporate, geographic, event and source namespace. The scope can be extended to other namespaces. The local namespaces can be linked to the central national namespace via identifiers. This is a modular and linked software, already realizing and modelling the structure used in the FOLIO complex: concepts of data model, flexible workflows, crowdsourcing, collaboration options, and authentication levels are used and realized. (Figure 5)

Authentication and access rights is a crucial component in each subsystem. Defining the domain (universe), institutions, branches, subbranches, groups, projects and individuals in a flexible way is essential, to allow a wide range of cooperation options and still define very precisely the workflows and the data available for the relevant partic-

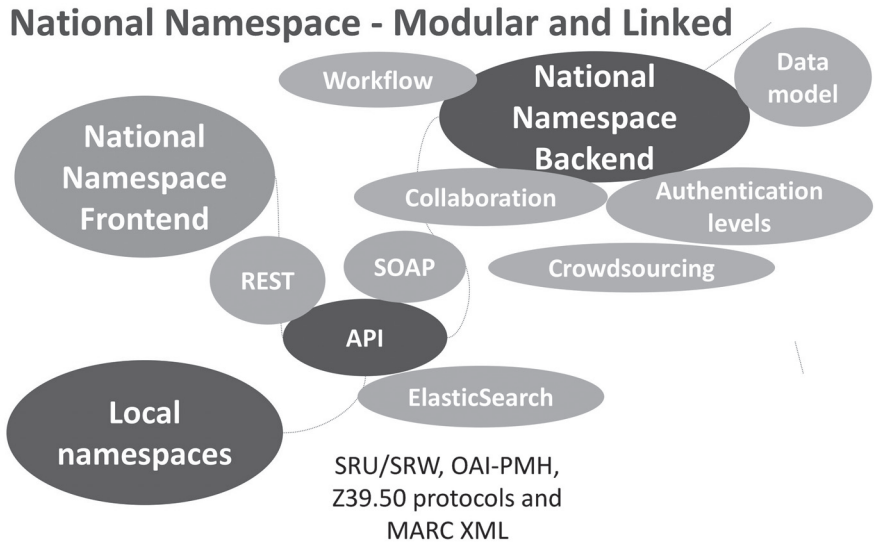


Figure 5.

ipants. In this context we ended up with a definition of a linked data relation between the participants, flexible, expandable, nonlinear description of options, and we derive the effective rights from those connections. There are options to grant or deny rights, by using this concept the definition of rights is more effective and effort-saving.

The national library by law has to enable free access to all available information. In this respect the national library has no right to select, what to collect and preserve and archive forever. This means, that the increase in data quantity has no end, and the amount of collected data is exponentially growing. The necessity of usable data models is even more imperative in the long term preservation concept. The definition of well-defined and centrally used entities, the connections between identified objects/entries, the usage of unique persistent identifiers all serve also the purpose of data minimization and saves storage.

The most significant transition regarding the data concept is, that with the introduction of the linked data, we rely on the data being there, where it should be. We do not duplicate the data – probably for the speedy access we cache some data temporarily, but we do not store them for long, we do not store unnecessary duplications. This makes data exchange unnecessary, so the significance of exchange format will disappear with time. This is the direction we are going, but we are all aware, that there is an ecosystem with the legacy data from the past (and present), which we have to feed for around

a decade from now on. This makes a complex interfacing necessary, between the linked universe and the existing systems based on data duplication.

And this is extremely important for long term preservation. We have to work on the concepts of structuring the data to be preserved long term, keeping track of incremental changes very carefully, and keeping storage and retrieval effort to an absolute necessary minimum. In this respect the aspects of work sharing and data models of linked data are fundamentally imperative.

The FOLIO community with a joint effort is able to define a solution for an efficient, cost saving and flexible long term preservation module. The National Széchényi Library is working on this now, exploring the new ways of a cloud based collaborative platform. Always keeping in mind that although what we think is the right and modern solution, tomorrow it might be an old fashioned one that could prevent further evolution. Therefore, whatever we do, the surrounding system must be flexible enough to embrace a new solution without bigger transitional pain. In order to keep libraries in the important position in the cultural landscape, we have to change almost everything of what we have in the present environment. In the future, we have to rely merely on constant change and evolution. In all definitions we form, we have to raise the abstraction level as high as we are able to do now – with the readiness to redefine everything tomorrow. And this is a constant contribution: the librarians are aiming for sustainability of the given and already achieved, the developers / strategists are looking for innovation and overcoming and replacing the existing structures by always better ones. This mutual support guarantees, that the current evolution in front of us is a gradual, step by step approach, sustaining the existing, at the same time realizing the future. We can trust in the wisdom of the broad community, which is always bigger than the wisdom of an individual, or of a same minded expert group. The world wide web provides the cradle of cooperation for of all us. We have to be brave enough to open up the doors widely for an extensive limitless cooperation, between domains and nations.

Further readings:
FOLIO community: www.folio.org
BIBFRAME model: www.bibframe.org
The National Széchényi Library: www.oszk.hu/en
The National Library Platform, webharvesting and digitization project:
http://www.oszk.hu/en/node/3490

National Library Platform Project (NLP/OKP) Parts of the Technical Description for the Procurement of the NLP System

The cited parts, which are mere examples out of our complex tender documentation, are illustrating the modularity, configurability and interoperability of the individual functions within our desired FOLIO system.

Library Administration (LMA)

An important requirement is that the system should enable a wide variety of configurations and parameters for the workflow processes, services and user interfaces handled by the various software components (modules). Our expectations of configurations in terms of the individual components are found in the list of requirements of the respective components.

The structural buildup of the library must be precisely mirrored so that a refined and granulated system of authorizations and privileges can be deployed. The structural hierarchy of the library must be maintained. Some parts of the hierarchy are to be registered in the CAD module as well: in order to operate the interlibrary loan services (ODR system) the registration of libraries of complex structure should be at least at four levels. In the case of libraries with more sites and locations and branch libraries each place of service must appear with its respective data (e.g., opening hours).

The LMA component handles the users' privileges/access rights. The privileges are hierarchic, which means the privilege of higher level is allowed to perform activities assigned to lower-level privileges. Apart from the levels of privilege built into the system there must be an option to define and assign new privileges centrally as well as per library. Every activity can be assigned to the users. It should also be possible to establish spheres of role: the individual privileges can be stuck together corresponding to the individual single user scope of work and at the same time allocated to users. The single user privileges can be restricted ensuing from the structural hierarchy of the library (e.g., they can borrow at certain places of service), or within certain workflow processes (e.g., in the course of cataloging those engaged in formal description have a different privilege from those doing subject cataloging meaning that specific fields are edited by either group. The user identifiers must be unique together with the library's identifier.

Circulation (CIR)

The rules of circulation and document distribution, the patron categories and the pertinent privileges, the settings related to circulation can be determined with the appropriate authorization: the number of loanable documents, duration of borrowing, renewal, sanctions ensuing delayed return (warnings, charge of delay, legal measures) hold requests, maintenance of opening and closing hours.

The creation of patron categories should be configurable: e.g., the library and/or library subunit of registration, the library/subunit concerning loan transactions, patron type, patron privilege, document type concerning loan transactions, maximum number of documents to be borrowed/placed on hold simultaneously, financial matters, etc. The limits to load/hold must be automatically tracked by the system and it mustn't allow exceeding the limits.

Patron identification should be at multiple level. It is possible to join national authentication systems. It should be enabled in the NLP system, depending upon configuration, to register patrons applying for the overall NLP system which enables the patron – relative to their privileges – to log in to several libraries with the same username. The central user authentication system assigns unique, unalterable identifier to the patron and by virtue of this the patron data can be registered in the individual libraries. The patron data can be altered by the authorized librarian.

Workflow Organization (WFL)

The expected functions of the WFL module fundamentally correspond to those of a universal workflow tool with the provision that the WFL should be able to cooperate with *all the substantial functions of the modules of NLP outlined herein*.

The WFL warrants tools for predefining workflow processes which can operate at all levels of the envisaged NLP model as well as between its levels. There must be an opportunity to conceive and run workflows the single stages of which are enacted outside of the NLP. The definition and setting up of the workflows must rely upon BPMN methodologies.

The WFL warrants workflow-monitoring support technically handled centrally yet logically matching the NLP's levels of data management in order for the currently initiated or automatically started processes to be traceable and controllable.

The scopes of the workflows can be defined at least by the division below:

- **Central processes** which are built upon common elements of the system and emphatically upon the common catalog and only those users are affected that possess privileges needed for central data management;
- Processes operating between the **central levels and a library** which typically should be operating in events when the termination or completion path of a process started at the central level is to be carried out at a given library's level;
- **Complex processes** when in the running of the process more than one library as well as central-level data management or process cycles are linked together. Such a process is interlibrary loan.
- Processes running **within one library** where each affected stage of process is enacted within one library or at most with the involvement of the library and one outside participant (see below).
- **Technical processes** that are carried on between informatics systems and/or services. Typically such processes are those of data export/import or other batch jobs running without user involvement.

In the WFL there is an option to set up event triggers. These triggers are tied to certain data recordings or user actions going on in the NLP modules, but it can also be an option to interpret it as a trigger for displaying the system contents (directories, files, etc.) defined outside of the NLP.

The relationship of predefined workflows and the triggers can be determined within the module. The result of the steps of a workflow must prevail as a trigger for the following step of the process. In the course of the process definitions the conditions of branch-off points can be determined and during supervision expected from the system it must be guaranteed that no inconsistent process definition can come about.

During the workflow definitions it is to be determined that the individual process steps are enacted by who or by what. For instance, in the course of handling a demand for procurement the recording of the demand can trigger a process of approval which process is made perhaps possible by more participants (approvers), and approval will come off when all the approvers have endorsed the demand. As the implementors (participants) of the process steps can be programs too, it should be possible to define processes where the agent of the process step is a particular program of informatics or web-based service.

The participants of each of the steps of the workflows can be classified into two major groups:

a) Users who play a role determined for the workflows, the types of users are not the same as the roles. The users by type are as follows:
   - NLP-level librarian users
   - Library-level librarian users
   - Users identified in other partner institutions
   - (Personified) named end users (patrons, researchers, etc.)
   - Anonymous users (users without sign-in accessing the system via the web)
   - NLP-level system administrators
   - Library-level system administrators
   - NLP-level technical users (external programs which effectuate their access to the system by means of personified user codes)

b) Technical agents non-displayable as a user. These can be the following:
   - Programs installed in local server environment running the NLP and prepared for running
   - Programs that can be called as a webservice and accessible with a permanent URL
   - Programs accessible as a webservice through a link resolver
   - RSS-based webservices

The above participants can be dynamically set up in the definition of the workflow. The participants belonging to group a) should appear in the system of privileges as authorized persons so that the system can concretely determine upon execution of the workflow the user who will be affected in the process step.

The workflows can be initiated by the user and not only by triggers. As far as the workflow is initiated by the user, then a special protagonist will emerge who is the STARTER (or initiator). In the course of the definition of the workflow the STARTER should be an exceptional role-player so that the steps which at all events are to be executed by it or they should reach it eventually can be identifiable. For example, when a request for acquisition is brought about by a librarian, then the librarian will be the STARTER, so when the process runs off and every approval has been granted then the STARTER will get notification that its demand has been approved. It follows, then, that during the running of the workflows the individual logical role players (who are linked via privileges) must appear in the system as actual role players.

During the definition, the WFL is capable of sending forward notices linked to any of the steps to role players determined at the logical level whose factual accessibility should be determined or resolved by the system at the actual running.

It is an option that during definition of the workflows the triggering of another workflow defined earlier can be fixed as a step.

The system must provide workflow simulation in the period of defining to test how the defined process will run off in practice.

The execution of concrete workflows is monitored by the system and, also, reports and queries built upon the logs should be available the running of which, like that of the definitions, is assigned to a special privilege.

General Expectations of Information Technology

The NLP software should support the data storage to be brought about in a (geo)redundant manner in terms of the handled contents, and the reserved operation, that is, in the event of service outage the redundant site is capable of taking over the functions in a short period of time and the recovery won't involve data loss or inconsistent state. The live service is invariably secured by a (primary) NLP installation running on a given locality, while there is a backup site installed in another locality with full functionality that can be run automatically or by manual intervention.

The system is of full modular structure. This allows for several basic configurations and the changing thereof:

– libraries participating in the system may use even a module divergent from the basic module, i.e., the system makes it possible to operate in parallel several versions of the same module,
  - the modules can be freely replaceable, their upgrade to more state-of-the-art versions is possible;
  - the modules can communicate with one another on standard interfaces, that is, the change of the module won't debilitate the operation of other modules;
  - the modules themselves are independent of the programming language, i.e., they can be written in any language.

The linguistic appearance of the modules (screens, images of printing) is freely selectable from among the languages offered, by installation and by user too. The num-

ber of languages available can be freely expanded, the system gives assistance to translations via user functions without programmers' interference.

The effectuation of the system is independent of the database manager. In the system, any database handling tool can be used the replacement of which is possible in the technical advancement.

As for the functions, in each function (catalog, treatment of privileges and users, institutional namespaces, etc.) several layers are to be materialized:

- common layer: participants handle common data on the commonly used software;
- central layer: data generated by software differing by institutions are synchronized and uploaded into the central data storage or downloaded from it by the systems;
- shared layer: the NLP system is capable of reaching data stored in the institutions' own systems without taking them over or harvesting them using them for view or listing. In this case both the application software and the data remain within the scope of the related institutions and won't merge with the NLP system.

In the course of data processing the catalogers have access to the data on the basis of levels of qualification and authentication. In a complex system they are enabled to move the data forward in the procedure, to raise them on higher level of qualification and authentication or relegate them on a lower level, to delegate tasks and assignments, to publish data, etc. The system warrants access to the data across the whole spectrum along this logic.

Display, User Interface, Facade (UIX)

The entire system will have a unified facade with the employment of a unified suite of icons. In addition, this component should ensure the appearance of an individual image for the libraries without programming/encoding, with parameters. On its own surface a member library should have the opportunity to feature the institution and its logo, configurable color schemes and fonts. On the emerging surface there should be available a virtual keyboard for the input of special characters.

There is an option to form an accessible user surface in compliance with level 'AA' of the W3C Accessibility Standard, e.g., a mode of operation of oversize contrast for the

visibility disabled, a loudspeaker help, the embedding of a text reciter, option of setting and enlarging font size.

Across the entire system there is a regularly updated context-sensitive help for the user surfaces. Also, the use of QR codes is possible (e.g., mobile help, display of information, a map of building for easy roaming). There are push-based notifications, configurable at the level of institution and by member library, on the events, useful information, contingent problems. There is an option to keep news, associated curiosities on perpetual display. The user surface must have an option for web-based update.

Access to the functions is possible according to privileges, in a configurable manner (at the institutional level and at the workflow level within it). The state of being logged-in should be finely visible on the opening page and, pursuant to the privileges of the signed-in user, the functions associated to them must come up (or get into active status). It is possible to use several functions in parallel, with suitability, and the navigation between the individual functions should be clearly viewable. While the function given in the user account is being used the other functions too can be viewable/accessible. The user can set the various elements of the display (opening of some functions on a new page, in a new window, a set of tools with parameters, etc.). Among data accessible within the functions there should be an option for query and sorting of hits.

In the NLP and Discovery system the display can be configurable. The data must be displayable concerning all types of documents, together with special data, and in every function configurable depending on the special treatment. It is possible to formulate brief and detailed forms of display for the bibliographic, authority and copy data.

In a number of functions of the NLP system (acquisitions, cataloging, circulation, etc.,) the formats of display should be configurably crafted and customizable according to the specifics and data of the tasks.

In the Discovery system it is possible to determine various formats (e.g., display of clusters of hits in formats with the determination of fields and field groups). In the patron settings too there must be an option to choose a format for saving and printing into the patron's library and for forwarding. The patron can choose from a list of predefined formats, or from a predefined list of data elements. The system is able to show indexed data originating from other databases via the standard API. There is an option for graphic display or data visualization on the basis of a list of hits (e.g., charts on the percent division of hits, the placement of data on a timeline).

When the digital contents are shown, it should be possible to create flexible visualizing pages(s) and to form various display surfaces according to document types and partial collections. There are to be customizable user surfaces: layers formed individually by partial collection (appearance, functions, helps, headers, footers).

The webpage displaying the document can be shared in the major social media pages and its link sent via email. There must be an option for displaying documents similar to the hit, flexibly defined on the basis of metadata.

Národní digitální archiv oslaví 3. narozeniny

Jiří Bernas, Národní archiv České republiky, Praha, ČR

Abstrakt
Prezentace i článek se budou věnovat řešení digitální archivace v ČR, zejména pak Národnímu digitálnímu archivu. Představena budou legislativní východiska, organizační a finanční zajištění digitálního archivu, principy jeho fungování a dalšího rozvoje. Zahrnuty budou dosavadní zkušenosti s akvizicí digitálních archiválií, s formáty a objemy dat.

Abstract
Presentation and article will deal with digital archiving in the Czech Republic, especially the National Digital Archive. Presented will be the legislative bases, organizational and financial provision of the digital archive, principles of its functioning and further development. Existing experience with the acquisition of digital archival records, formats and data volumes will be included.

Úvod

První úvahy o možnostech digitální archivace se v českém archivnictví objevily přibližně v polovině devadesátých let 20. století. Na původní myšlenky o nutnosti řešit ukládání digitálních dokumentů navázaly konkrétní výzkumné projekty realizované za účasti Odboru archivní správy MV, ČVUT a Národního archivu v letech 2001-2005, na jejichž základě dostaly záměry konkrétnější podobu.

Snaha o řešení archivace digitálních archiválií vyústila v usnesení vlády č. 11 ze 7. ledna 2004. Tímto usnesením uložila vláda svému místopředsedovi a ministru vnitra zpracovat ve spolupráci s ministrem informatiky projekt dlouhodobého uchovávání a zpřístupňování dokumentů v digitální podobě. V předkládací zprávě usnesení byla nejen konstatována a zdůvodněna nutnost řešit dlouhodobé uchovávání digitálních dokumentů, ale rovněž byl určen postup směřující k vybudování digitálního archivu.

Nejprve měl být při Národním archivu sestaven tým, jehož úkolem byla příprava podkladů pro zpracování projektu. Na jeho základě pak měl být vybudován digitální archiv.

Naléhavost řešení problému digitální archivace byla následně podpořena i rozvojem e-governmentu, který není bez vyřešení digitální archivace v zásadě možný.

Na základě uvedeného vládního usnesení byl v Národním archivu ustaven na konci roku 2005 ustaven tým, jehož úkolem bylo shromáždit podklady a vypracovat zadávací dokumentaci pro výběr zpracovatele projektu digitální archiv. Vybraný dodavatel pak v období od července 2007 do února 2008 zpracoval technologický projekt Pracoviště pro dlouhodobé uchovávání a zpřístupňování dokumentů v digitální podobě (http://www.nacr.cz/zpravy/projekt_nda.aspx). Na jeho základě byly následně podniknuty kroky k vybudování specializovaného pracoviště zabývající se dlouhodobým uchováváním digitálních dokumentů při Národním archivu. Jedná se o pracoviště, pro které se vžilo pojmenování Národní digitální archiv.

Koncepce péče o digitální archiválie v ČR

Digitální dokumenty plynule navazují v produkci původců na tradiční dokumenty. Digitální archiválie jsou proto chápány jako součást stávajících archivních fondů, ve kterých rovněž plynule navazují na tradiční archiválie. Oba typy archiválií se liší jen svou fyzickou podstatou a formou, jakou je třeba o ně pečovat. Na základě tohoto přístupu je digitální archiv chápán jako specializované pracoviště zajišťující dlouhodobé uložení a čitelnost digitálních archiválií. Tyto archiválie však zůstávají v archivní péči příslušných archivů, které provádějí jejich výběr a řeší jejich archivní zpracování.

Problematika dlouhodobého uchovávání digitálních dokumentů je však finančně náročná. Proto bylo zvoleno řešení, jímž je jeden hlavní digitální archiv při Národním archivu označovaný jako Národní digitální archive. Národní digitální archiv je tedy v České republice jediným digitálním archivem pro uložení digitálních archiválií v péči Národního archivu, Archivu bezpečnostních složek a státních oblastních archivů. Ostatní akreditované archivy mají možnost ukládat své digitální archiválie v Národním digitálním archivu, vybudovat si vlastní digitální archiv nebo ukládat v digitálním archivu jiného archivu. Pokud si archiv vybuduje vlastní digitální archiv je sou-

částí udělení oprávnění ukládat digitální archiválie i potvrzení o úspěšném přenosu jimi spravovaných archiválií do Národního digitálního archivu. Každý digitální archiv by tak měl být schopen v případě svého zániku předat spravované digitální archiválie do Národního digitálního archivu.

Legislativní východiska

Základ archivní legislativy tvoří zákon č. 499/2004 Sb., o archivnictví a spisové službě. V oblasti péče o digitální dokumenty tento zákon určuje kompetence ministerstva a archivů, stanovuje podmínky pro udělení či odnětí oprávnění k ukládání archiválií v digitální podobě a v neposlední řadě definuje úkoly portálů pro zpřístupnění archiválií v digitální podobě.

Ustanovení zákona dále upřesňují následující prováděcí předpisy: vyhláška č. 645/2004 Sb., kterou se provádějí některá ustanovení zákona o archivnictví a spisové službě, vyhláška č. 259/2012 Sb., o podrobnostech výkonu spisové služby, a oznámení Ministerstva vnitra, kterým se zveřejňuje Národní standard pro elektronické systémy spisové služby (VMV čá 57/2017).

Vyhláška č. 645/2004 Sb. v oblasti digitálních archiválií definuje zejména minimální množinu popisných metadat. Pro péči o digitální dokumenty a digitální archiválie jsou významnější zbylé dva prováděcí předpisy. Vyhláška č. 259/2012 Sb. Stanovuje kromě povinností při příjmu, správě a vyřazování digitálních dokumentů výstupní datové formáty, tj formáty pro výstup z elektronického systému spisové služby, formáty pro dokumenty ukládané ve spisovně a pro dokumenty předávané do digitálního archivu. Národní standard pro elektronické systémy spisové služby pak stanovuje vlastnosti, které elektronický systém spisové služby musí mít a podobu SIP, tj. datových jednotek s logicky definovanou strukturou pro předání dat a jejich metadat za účelem jejich dlouhodobého uložení.

Organizační zajištění

Národní digitální archiv není samostatnou institucí, ale tvoří nedílnou součást Národního archivu. Jedná se vlastně o populární označení pracoviště zabývající se dlouhodobým uchováváním dokumentů v digitální podobě. Tým spravující a rozvíjející Národ-

ní digitální archiv je součástí Oddělení kontroly výkonu spisové služby, fondů státní správy po roce 1992 a elektronických dokumentů Národního archivu. Konkrétně se jedná o úsek ICT, který zajišťuje vlastní provoz technologií, vlastní vývoj aplikací, kontrolu aplikací vyvíjených externím subjektem, testování a rozvoj systémů, a úsek metodiky, který se podílí na definici požadavků, kontrole jejich naplnění a přípravě legislativy a metodik péče o digitální dokumenty.

Tým spravující Národní digitální archiv v současné době tvoří čtyři vývojoví pracovníci ICT (programátoři, analytici), tři operátoři ICT, čtyři správci ICT a tři metodici.

Finanční zajištění

Stabilní financování je jedním ze základních předpokladů dlouhodobé archivace digitálních dokumentů. V ČR se v zásadě tento předpoklad podařilo úspěšně naplnit, neboť nedlouho po dokončení výše zmíněného projektu Pracoviště pro dlouhodobé uchovávání a zpřístupňování dokumentů v digitální podobě, tj. v dubnu 2008, schválila vláda svým usnesením č. 447 z roku 2008, k zabezpečení plnění úkolů ve věci vybudování Národního digitálního archivu, čerpání finančních prostředků na vybudování Národního digitálního archivu. Později vláda usnesením č. 611 z 9. srpna 2013, k Informaci o stavu projektu Národního digitálního archivu a předpokladech zajištění jeho standardního provozu, uložila ministru vnitra zabezpečit finanční prostředky nezbytné pro provoz a rozvoj Národního digitálního archivu. Zmiňované finanční prostředky zahrnují investiční (rozvoj), provozní i mzdové náklady a počínaje 2. pololetím roku 2014 byly přidány do rozpočtu Národního archivu.

Původním záměrem bylo vybudovat Národní digitální archiv jako ucelený systém dodaný externím dodavatelem. Za tímto účelem byl zpracován projekt Národní digitální archiv spolufinancovaný ze Strukturálních fondů Evropské unie. Projekt zahrnoval vybudování prostor pro hlavní i záložní úložiště a vývoj informačního systému. Byl zahájen v roce 2011, nicméně pro potíže s realizací veřejné zakázky jeho hlavní části, informačního systému, musel být v roce 2014 ukončen. Vzhledem ke komplikacím bylo již v průběhu projektu zvažováno náhradní řešení. Po zrušení projektu pak byl Národní digitální archiv vybudován jako modulární systém založený na LTP Archivematica. Tento systém je v současnosti modernizován opět z prostředků Strukturálních fondů EU v rámci probíhajícího projektu Národní digitální archiv II.

Popis NDA

Národní digitální archiv je koncipován jako modulární systém a vychází ze standardu OAIS (ISO 14721:2003 – Open Archival Information System). Dokument a jeho metadata tvoří balíček s jednotnou strukturou. Podle standardu OAIS jsou tyto balíčky nazývány SIP – Submission Information Package (balíček přijímaný do digitálního archivu), AIP – Archival Information Package (balíček pro uchování v digitálním archivu) a DIP – Dissemination Information Package (balíček poskytovaný digitálním archivem). Národní digitální archiv tvoří dva informační systémy IS NDA a IS Archivní portál. Oba systémy fungují samostatně a spolu komunikují omezenou množinou protokolů.

Základem IS NDA je LTP Archivematica, který je doplněn o moduly umožňující příjem SIP balíčků definovaných Národním standardem pro elektronické systémy a SIP pro výběr mimo skartační řízení, distribuci AIP do záložního pracoviště a přístup k AIP a metadatům. Systém je uzavřený a umístěný v samostatném segmentu sítě. Běžnému uživateli je zcela nedostupný.

IS Archivní portál je určen pro komunikaci Národního digitálního archivu s vnějšími uživateli. Jedná se o webovou aplikaci, která umožňuje zejména provádět výběr archiválií ve skartačním řízení a mimo něj. Ostatní funkce jsou momentálně implementovány v minimálním rozsahu a jejich rozvoj je plánován až v rámci projektu Národní digitální archiv II.

Způsob uchování

Dlouhodobé uchování digitálních dokumentů sestává ze zachování dokumentu a ze zajištění jeho čitelnosti. Problematika zachování dokumentu je řešena vícenásobným uložením dokumentu. Dokument je uložen souběžně do hlavního a do záložního úložiště, v nichž je uchováván a je kontrolováno, zda jsou totožné. Neprobíhá tedy synchronizace úložišť, ale úložiště jsou na sobě nezávislá. V současné době jsou data uložena na diskových polích umístěných v budovách Národního archivu v Praze 4 a v Praze 6. V následujícím roce plánujeme přesun záložního úložiště do Hluboké nad Vltavou.

Čitelnost dokumentu je zajišťována pomocí migrace. Pozornost je prioritně věnována formátům přicházejícím do digitálního archivu od původců, zejména úřadů. Vyhláška

č. 259/2012 Sb. definuje výstupní formáty pro statické textových dokumentů a statické kombinované textové a obrazové dokumenty (PDF/A, ISO 19005), statické obrazové dokumenty (PNG, JPG, TIF revize 6), dynamické obrazové dokumenty (MPEG-1, MPEG-2, GIF), zvukové dokumenty (MP2, MP3, WAV-PCM), databáze (XML s dostupným DTD nebo XSD) a metadata dokumentů ve spisové službě (XML definované Národním standardem, přílohou č. 2).

Při ukládání do LTP Archivematica jsou formáty souborů identifikovány a migrovány podle defaultních pravidel systému. V současné době se Národní digitální archiv soustřeďuje zejména na akvizici digitálních dokumentů, a proto je podrobnější identifikace formátů a stanovení vlastních migračních strategií úkolem, který teprve čeká na řešení. Aktuálně je zkoumáno využití formátu SIARD pro ukládání databází.

Jak již bylo zmíněno, soustředí se Národní digitální archiv zejména na akvizici. Problematika přístupu k uloženým digitálním archiváliím je proto řešena až nyní v rámci projektu Národní digitální archiv II. V jeho rámci bude Archivní portál rozšířen o modul eZpřístupnění, který umožní uživatelům přístup k digitálním archiváliím. Přistupovat bude možno anonymně (půjde-li o archiválie přístupné bez omezení) nebo na základě badatelského listu.

Akvizice

K akvizici archiválií dochází dvěma způsoby: výběrem ve skartačním řízení a výběrem mimo skartační řízení. Výběr probíhá prostřednictvím Archivního portálu. Je důležité zdůraznit, že výběr ve skartačním řízení se týká všech dokumentů evidovaných v elektronickém systému spisové služby. Tedy i analogové dokumenty evidované v elektronickém systému procházejí výběrem prostřednictvím Archivního portálu.

Výběr ve skartačním řízení

Výběr archiválií ve skartačním řízení zahajuje zpravidla původce. Po přihlášení k webovému rozhraní archivního portálu založí? skartační řízení a nahraje SIP. Přitom není nutné nahrávat SIP jednotlivě, ale je možné, a vzhledem k počtu SIP i výhodné, balíčky před nahráváním zabalit metodou ZIP a nahrávat je jako jeden soubor.

Po nahrání jsou jednotlivé SIP podrobeny antivirové kontrole a validaci. Podoba SIP je stanovena Národním standardem pro elektronické systémy spisové služby a zde nastávají první problémy. Ačkoli první verze Národního standardu je již z roku 2009 (novelizované verze z roku 2012 a 2018) mají mnohé systémy problém s jeho vytvořením. Problémem se ukázal již požadavek, aby vzniklé XML odpovídalo předepsanému schématu. Národní archiv proto na svých stránkách zpřístupnil validátor SIP, který umožňuje původci či dodavateli systému spisové služby zkontrolovat si jím produkované SIP. Postupně se však situace zlepšila a v dnešní době výskyt chybně konstruovaných SIP poklesl na minimum. Vyvstal však nový problém, a to v podobě kvality vyplňovaných metadat. Do značné míry se zde projevují nedostatky ve vedení spisové služby, které nebyly do této doby viditelné. Některé údaje nejsou vyplňovány vůbec, některé nedbale. Při příležitosti novely Národního standardu tak došlo i k doplnění validátoru SIP o testy obsahu a jeho konzistence.

SIP pro analogový dokument či spis tvoří pouze metadata, ačkoli není vyloučeno, aby byl v balíčku přiložen i koncept dokumentu v digitální podobě. SIP digitálního dokumentu nebo spisu sestává z metadat a komponent. V případě skartačního řízení připouští vyhláška č. 259/2012 Sb. použití redukovaného SIP tvořeného pouze metadaty. V takovém případě je plnohodnotný SIP předkládám až na výzvu archiváře. Možnost překládat redukovaný SIP digitálního spisu nebo dokumentu byla zavedena, aby se snížilo množství dat přenášených při skartačním řízení.

Po nahrání všech balíčků předává původce řízení příslušnému archivu a získává seznam všech nahraných SIP s časem nahrání, velikostí balíčku a kontrolním součtem SHA512. Seznam poté odesílá tradiční cestou příslušnému archivu jako přílohu návrhu na výběr ve skartačním řízení.

Archivář se po obdržení návrhu přihlásí k webovému rozhraní Archivního portálu, kde si vyhledá příslušné řízení a prostřednictvím modulu eSkartace provede výběr. Pokud jsou posuzovány digitální dokumenty, které původce zaslal v podobě redukovaného SIP (pouze s metadaty), může si vyžádat předložení plnohodnotného SIP. Tento požadavek zasílá původci v podobě XML souboru, jehož podoba je určena Národním standardem. Původce pak na výzvu nahraje požadované dokumenty opět prostřednictvím Archivního portálu.

Po ukončení výběru je vygenerován XML soubor obsahující rozhodnutí archiváře: vybrat za archiválii, zničit nebo vyřadit ze skartačního řízení. Podoba XML je definována Národním standardem. Pokud nejsou při skartačním řízení vybrány archiválie, původce vyznačí ve svém systému spisové služby u každého dílu, spisu či dokumen-

tu informace o skartačním řízení a může ze systému odstranit komponenty. V opačném případě vyčká s vyznačením do úspěšného uložení vybraných archiválií v digitálním archivu a připraví plnohodnotné SIP vybraných dílů, spisů a dokumentů k přejímce.

V současné době probíhá v rámci projektu Národní digitální archiv II modernizace, která odstraní nutnost nahrávat SIP přes webové rozhraní a přinese možnost zasílat SIP prostřednictvím webové služby přímo ze systému spisové služby původce. Analogicky bude probíhat případné předložení plnohodnotného SIP a zaslání rozhodnutí o výběru.

Výběr mimo skartační řízení

Výběr mimo skartační řízení je určen zejména pro digitální dokumenty, které nejsou evidovány v elektronickém systému spisové služby. Typicky se může jednat o dokumenty soukromé osoby strukturované pouze za pomoci adresářové struktury na disku.

Při výběru mimo skartační řízení zahajuje proces archivář a po zvážení okolností pak nahrává soubory, které jsou předmětem výběru, prostřednictvím webového rozhraní Archivního portálu sám nebo vytvoří pro původce dočasný účet. Při založení výběru mimo skartační řízení zadá archivář základní informace k němu: číslo jednací, původce, archiv. Adresářová struktura se vždy nahrává zabalená metodou ZIP. Nahrané soubory projdou po nahrání antivirovou kontrolou.

Archivář následně provádí výběr ve webovém rozhraním připomínajícím správce souborů typu Norton Commander. V levém okně vidí nahranou adresářovou strukturu, v pravém okně pak vytváří strukturu datasetů ze souborů vybraných za archiválii. K jednotlivým datasetům, adresářům i souborům následně vyplňuje metadata. Některá z metadat se aplikace snaží ze souborů získat automaticky (EXIF, systémové informace).

Po ukončení výběru vytvoří aplikace z definovaných datasetů SIP, které odešle k přejímce. Podoba SIP z výběru mimo skartační řízení není ukotvena legislativně, ale jedná se o interní formát Národního archivu. Zkušenosti z jeho tvorby Národní archiv následně využil ve svých návrzích při přípravě SIP pro výběr ve skartačním řízení.

Přejímka

V rozhraní archivního portálu je archiváři k dispozici seznam připravených, probíhajících či ukončených přejímek.

Ke každé přejímce je nutno vyplnit základní metadata: číslo jednací, přírůstkové číslo, pečující archiv, archivní fond a původce. Údaje, které jsou již systému známy, jsou doplněny automaticky. Nahrávání SIP k přejímce probíhá obdobně jako při skartačním řízení.

Zpracovávaná přejímka je nejprve normalizována v modulu ePřejímky. tento modul je určitým reliktem. Do nedávné doby nebyla struktura SIP pro skartační řízení v Národním standardu jednoznačně definována, proto z různých SIPů přicházely mírně odlišné SIP. Dnes je již tento SIP definován jako adresářová struktura obsahující soubor mets.xml a adresář komponenty, která může být zabalena metodou ZIP, ale dříve se bylo možno setkat při přejímce ze skartačního řízení s různými SIP. Nebylo například ustáleno pojmenování samotného souboru XML s metadaty dle Národního standardu. Situaci si znesnadnil i Národní archiv sám, když s ohledem na možné komplikace při zpracování XML obsahujícími komponenty (původní verze Národního standardu předpokládala komponenty přímo v XML) připouštěl i SIP v podobě adresářové struktury s komponentami mimo XML zabalené metodou SIP.

Modulu ePřejímky jsou SIP zasílány prostřednictvím WSDL a tento modul zároveň představuje hranici mezi Archivním portálem a IS NDA. Přejímka je následně zpracovávána modulem Příjem, od kterého je oddělen firewallem a k němuž nemá přístup.

Modul Příjem si přejímku stáhne za pomoci SFTP a ověří pomocí kontrolních součtů, že přenos proběhl v pořádku. Součástí přejímky je i XML s rozhodnutím archiváře, podle kterého Příjem zkontroluje, zda byly předány všechny díly, spisy, dokumenty a datasety. Pokud nebyly předány kompletně nebo bylo předáno něco navíc, je přejímka zamítnuta. Pokud je kontrola přejímky úspěšná, jsou jednotlivé SIP transformovány do podoby vhodné pro přijetí modulem Archivematica. Součástí této transformace je přidání souboru MCPprocessing.xml, ve kterém je definován postup dalšího zpracování v LTP Archivematica. Jednotlivé SIP jsou následně předány tomuto modulu ke zpracování. Výsledek je zachycen v XML souboru, který je předáván zpět modulu Příjem a též modulu Distribuce. Modul Příjem vyčká na zpracování celé přejímky a poté na základě výsledků zpracuje výsledný seznam uložených dílů, spisů, dokumentů nebo datasetů, který cestou ePřejímek předá zpět Archivnímu portálu. Podoba seznamu je určena Národním standardem. Tento seznam zasílá archivář tradiční cestou původci.

Jde-li o výběr ve skartačním řízení, nahraje původce výsledný seznam do svého systému spisové služby, který vyznačí, že daný díl, spis či dokument je uložen v digitálním archivu a zaznamená rovněž identifikátor digitálního archivu, pod nímž ho lze vyhledat. Modul Distribuce zajistí na základě obdrženého XML jeho uložení v záložním úložišti.

Objemy dat

Určení objemů dat, které lze pro uložení v digitálním archivu očekávat je značně problematické. Národní archiv sice provedl osobní i dotazníkové šetření původců, ale jejich hlavním záměrem bylo, že údaje jsou velmi roztříštěné. Zdá se, že současné systémy spisových služeb nejsou schopny poskytnout informaci, kolik bytů dat spravují a kolik z nich je označeno skartačními znaky S, V nebo A. Obdobně bylo při osobních průzkumech obtížně zjišťováno, kolik dat původce spravuje a kolik z nich by se mohlo v budoucnu stát archiváliemi. Zpravidla bylo možno získat informaci, kolik diskového prostoru data zabírají, horší bylo zjistit, kolik zabírají zálohy a duplicity. Objemy dat pro digitální archiv proto byly odhadovány na základě zkušeností s analogovými dokumenty: počty jednotlivých typů dokumentů krát jejich odhadovaná velikost. Blíže je tento výpočet popsán v Projektu pracoviště pro dlouhodobé uchovávání a zpřístupňování dokumentů v digitální podobě.

V Národním digitálním archivu je v současné době uloženo 6 GB dat. Mimo ně Národní archiv spravuje ještě 5,24 TB dat převzatých před spuštěním Národního digitálního archivu.

Závěr

Národní digitální archiv se řadí k těm digitálním archivům, které vznikly na základě vize vyřešit problém digitální archivace jedním projektem. Vinou komplikací při realizaci projektu se transformoval do podoby menšího, postupně se rozvíjejícího systému. Do vínku mu však při jeho spuštění (od 1. 11. 2014 zkušební provoz, od 1. 1. 2016 rutinní) bylo dáno stabilní financování a stabilní projektový tým, což je z dlouhodobého hlediska nejdůležitější.

Udržitelnost NDK, rozvoj digitalizačního pracoviště NK a MZK a jejich vztah k dalším knihovnám

Petr Kukač, Národní knihovna České republiky, Praha, ČR

Abstrakt
NDK znamená Národní digitální knihovna. Jedná se o systém pro digitalizaci, dlouhodobé uchování a zpřístupnění moderních bohemikálních tisků a také pro webarchív českých domén, od sklízení až po dlouhodobé uložení. Systém byl vyvinutý a implementovaný v letech 2010 – 2014 s přispěním dotace EU. Je tvořen centrálním výpočtovým a ukládacím pracovištěm v Praze, dvěma digitalizačními linkami po jedné v NK ČR v Praze a v MZK v Brně a pracovištěm pro správu webarchívu v Praze. Systém je v trvalém provozu produkujícím nová data od srpna 2013 s výhledem jeho dalšího dlouhodobého provozování. Petr Kukač, ředitel odboru digitalizace NK ČR a vedoucí pracovního týmu NDK představí úskalí zajišťování každodenního provozu v podmínkách omezeného financování ze státního rozpočtu při nutnosti zachovat aktuálnost prostředí NDK v dynamicky se rozvíjejícím světě digitálního knihovnictví.

Abstract
NDK means the National Digital Library. It is a system for digitization, long-term preservation and access to modern bohemian prints as well as for web archives of Czech domains, from harvesting to long-term storage. The system was developed and implemented between 2010 and 2014 with the contribution of the EU subsidy. It consists of a central computing and storage workplace in Prague, two digitization lines one in the National Library of the Czech Republic in Prague and in the Moravian Library in Brno and a workplace for the administration of the web archive in Prague. The system is in constant operation producing new data from August 2013 with a view to its further long- term operation. Petr Kukač, Director of the Digitization Department of the National Library of the Czech Republic and the head of the NDK work team, will present difficulties in ensuring day-to-day operation under conditions of limited fund-

ing from the state budget, while maintaining the up-to-date NDK environment in the dynamically evolving world of digital librarianship.

Národní digitální knihovna. Slovní spojení, které ne vždy je použito ve významu, jak jej vnímá Národní knihovna České republiky. Je proto na místě správný výklad pojmu Národní digitální knihovna osvětlit.

Již někdy v roce 2008 vznikla myšlenka na vybudování komplexního prostředí, které umožní masovou digitalizaci moderní literatury. Tak, aby knihovna byla schopná poskytovat čtenářům své fondy v digitální podobě v reálně použitelné formě a hlavně měřítku. Protože předpokládaná nebo odhadovaná finanční náročnost realizace takové myšlenky mnohonásobně převyšovala rozpočtové možnosti Národní knihovny, bylo rozhodnuto přihlásit aktivitu jako projektovou přihlášku do Integrovaného Operačního Programu. Podle specifických podmínek IOP byl původní velmi ambiciózní záměr zredukován do podoby, jak prošel procedurou schvalování dotace a jak vypadá prakticky dodnes, s výjimkou drobných dílčích změn.

Projekt byl navržen v partnerství Nositele projektu – Národní knihovny ČR a Partnera projektu – Moravské zemské knihovny. Cílem projektu NDK, jak byl tehdy konstituován, bylo vytvoření fungujícího systému „Národní digitální knihovny“ jako součásti vznikající České digitální knihovny a Europeany. Tohoto cíle mělo být dosaženo prostřednictvím následujících výstupů projektu:

- Vybudovaná potřebná technická infrastruktura
- 2 fungující digitalizační pracoviště v Praze a v Brně
- Důvěryhodný repozitář pro dlouhodobé uloženi digitálních dokumentů
- Systém zpřístupnění digitálních dokumentů
- Integrovaný systém digitalizace, uložení a zpřístupnění knihovních dokumentů propojený s významnými národními a evropskými portály
- Digitalizace, dlouhodobé uložení a zpřístupnění 26 000 000 stran digitalizovaných dokumentů
- Harvesting, dlouhodobé uložení a zpřístupnění 4 000 000 000 webových souborů.

Celková plánovaná doba realizace projektu byla naplánována na 57 měsíců (duben 2010 – prosinec 2014). Celkový rozpočet projektu byl vykalkulován na bezmála 300 milionů Kč. V souladu s dotačními pravidly IOP 85 % nákladů bylo hrazeno ze Strukturálních fondů EU prostřednictvím programu IOP, zbylých 15% z prostředků Národní knihovny jako žadatele.

Za oněch 57 měsících se nám podařilo vyvinout komplexní prostředí složené z ICT infrastruktury, speciálních knižních skenerů, softwarového řešení, jednotlivých pracovišť včetně pracovních míst a souboru procesů, postupů a znalostí. Nejen tedy vyvinout, ale také nasadit, implementovat, otestovat v pilotním provozu a nakonec převést do plně produkčního režimu. Vlastně to všechno muselo být dokončeno ne v definovaných 57 měsících, ale o rok dříve. Aby byl dostatek času pro digitalizaci 26 milionů stran nejpozději do konce roku 2014 a tím splnit jeden z hlavních ukazatelů úspěšného ukončení projektu.

Prakticky veškeré výpočetní operace a ukládání provozních i výsledných dat je soustředěno do centrálního pracoviště v Národní knihovně v Praze, konkrétně v Centrálním depozitáři v Praze – Hostivaři. Do tohoto centra jsou propojena všechna pracoviště tvořící technologickou digitalizační linku. Jedna digitalizační linka běží ve stejném objektu v Hostivaři, prakticky identická pak v Moravské zemské knihovně v Brně. Tyto digitalizační linky produkují datové balíčky připravené k uložení do repozitáře dlouhodobého zabezpečeného úložiště a do zpřístupňovacího systému. Toto uložení a všechny s tím související činnosti jsou realizovány opět na jednom bodu v Praze, kde také LTP systém běží. Vlastní dlouhodobé uložení dat je řešeno technologií páskových knihoven a magnetických pásek.

Aby takto implementovaný systém mohl dlouhodobě fungovat, je samozřejmě nutné trvale aktivně udržovat. Už v Studii proveditelnosti, povinné příloze projektové žádosti na IOP, jsou poměrně zevrubně definovány činnosti, úkony a náklady, které by v ideálním případě měly být realizovány proto, aby systém Národní digitální knihovny nejen splnil dotační podmínky vyžadující udržet výsledky projektu nejméně 5 let po jeho ukončení, ale aby na konci této doby, konkrétně po 31. prosinci 2019, byl systém v kondici umožňující pokračovat v nezměněné intenzitě provozu po další léta. Dalšími léty je ideálně myšlena doba, dokud nebude realizován původní záměr, totiž zdigitalizovat, dlouhodobě uložit a zpřístupnit veškeré kulturní dědictví tvořené moderními bohemikálními fondy spravovanými v Národní knihovně. Pro uvedení do reality, jde o nejméně 310 milionů stran. Dnes po 5 letech produkce je zdigitalizováno a zpřístupněno přibližně 52 milionů přepočtených stran A4. Čistě aritmeticky to znamená, že dosáhnout původního záměru by se mohlo podařit po nejméně 30 letech dalšího provozu, ale bude to nejspíš daleko déle

Tak jako mnoho jiných aktivit, které se zahajují nově, se doporučuje začít od jednodušího a postupovat směrem k složitějšímu. V prostředí Národní digitální knihovny to znamená, že digitalizovány byly nejdříve novodobé monografie menších formátů ve velmi dobrém fyzickém stavu; ty byly skenovány na automatických a robotických ske-

nerech. Až později se v daleko větší míře začaly uplatňovat svázané ročníky periodik, velkoformátové předlohy a také svazky, u kterých fyzický stav dovolí jen velmi šetrné ruční zpracování. Tento trend se stále zrychluje. Také proto se později dokoupily ještě další ruční skenery, jmenovitě i2S Suprascan A1 HD, a 4DB scanVpage.

Každá komponenta dlouhodobě provozovaného ICT systému má svou technickou i morální životnost. Nejinak je tomu i u Národní digitální knihovny. Standardní hardware sestávající z pracovních stanic, serverů, diskových polí, síťových prvků a podobně je v drtivé většině používán původní, k obnově zatím nedošlo. Jedinou výjimkou je provozní diskové pole, kde došlo k havárii řadičů, data se ale podařilo zachránit. Speciální hardware, typicky knižní skenery, jsou také provozovány od počátku, obnovit bylo nutné zatím jen ploché skenery Plustek A300, které už byly skutečně opotřebené a vykazovaly sníženou spolehlivost i kvalitu výstupů. Generační obnova hardware stále čeká na svou realizaci, předpokladem je, že proběhne v příštím roce.

Digitalizace knihovnictví je obor, který v poslední době zažívá velký boom. S tím chtě nechtě souvisí zavádění nových standardů, doporučení, formátů souborů, a podobně. Vedle knihovnictví, také ICT se neustále vyvíjí, jsou uváděny na trh nové technologie, operační systémy, JAVA, atd. a ty staré se dostávají mimo podporu. Všechny tyto aspekty Národní digitální knihovnu ovlivňují.

V prostředí evropské legislativy chránící volný trh a zamezující neprůhledným finančním tokům, jinými slovy v podmínkách zákona č. 134/2016 Sb. o zadávání veřejných zakázek, je dodavatelské poskytování služeb velmi složité zajistit. De facto jsme stále ve zpoždění za technickým vývojem, na který sice můžeme zareagovat v jednotlivých případech nutnosti změny včas, ale než se zrealizuje zadávací řízení, vybere dodavatel, uzavře smlouva a provede realizace změnového požadavku, často uplyne hodně času. To vše za předpokladu, že jde o změny, které se nevynutí zásadní přepracování celého prostředí Národní digitální knihovny. Takto se nám podařilo implementovat do systému například nové specifikace pro metadatové záznamy monografií a periodik – verzi z roku 2015. Nejnovější metadatové specifikace, platné od letošního roku, teprve implementujeme, práce budou dokončeny v průběhu října 2018. To je velmi důležité pro vzájemnou výměnu souborů s jinými knihovnami a možnost zpracování takto importovaných titulů. V České republice totiž funguje dotační program na podporu digitalizace, jehož jednou z podmínek je, že příjemce dotace musí výstupy své práce předat do Národní knihovny k uložení v LTP.

Průběžně jsou také nasazovány nové verze zpřístupňovacího systému Kramerius 5. Pro příští rok ale bude nutné přejít na výrazně přepracovanou verzi 6, bude to zřejmě

spojené s dodávkou nového hardwaru, samotná migrace dat pak bude trvat zřejmě několik měsíců.

Co se zatím nepovedlo, je provedení generační obměny základního frameworku, na kterém celé prostředí Národní digitální knihovny běží. Jde o produkt české firmy AiP Safe s.r.o., svým způsobem jedinečný. Odtud také pramení potíže při realizaci zadávacího řízení. Je to pro celou Národní knihovnu velká výzva, protože bez rychlé obnovy přijdeme o podporu výrobce.

Národní digitální knihovna používá v průběhu zpracování obrazů nekomprimovaný souborový formát TIFF. Poplatně době vzniku systému, TIFF verze 5. Dnes aktuální verze 6.0 přidala některé nové vlastnosti, na které technologická linka nebyla připravena, a do dnešního dne se nám nepodařilo ji připravit. Namísto toho musíme všechny obrazy, které získáme v jiné verzi TIFFu než 5, do této verze konvertovat. V případě přechodu z vyšší verze jistě nesystémové řešení. Obrazy ve formátu TIFF přitom do zpracování proudí nejen z třetích stran, z knihoven, které sami digitalizují a své produkty do Národní knihovny posílají, ale také z nových skenerů, které jsme si sami pořídili a jejichž obslužný software je prostě na rozdíl od NDK s dobou. Naštěstí náš dodavatel skenerů je velmi chápavý a vstřícný a významně nám při hledání řešení pomáhá.

Tak, jak komplex Národní digitální knihovny pro určité operace s obrazy nebo daty používá jednoúčelové aplikace a utility často vytvořených na bázi open source řešení, dostali jsme se již také zde do situace, že vývojářská komunita konkrétního programu zanikla a s ní i dostupnost jakékoli podpory, nových verzí a podobně. Konkrétně např. program ScanTailor, využívaný třeba pro ořezy obrazů. Poslední dostupné verze jsou stále 32-bitové a neumožnují zpracování velkých souborů, které u nás vznikají třeba digitalizací velkých formátů mapových děl. V současnosti proto hledáme jeho náhradu a způsob výměny v lince.

Ukončení podpory ze strany výrobce se netýká jen softwarů. Jedním z klíčových subsystémů Národní digitální knihovny je IBM Information Archive zajišťující přípravu dat pro ukládání na magnetické pásky a celou přidruženo agendu. I tento systém je již nejméně rok mimo podporu výrobce, přesto u nás stále běží a náhrada zatím nebyla dořešena.

Bez obnovy či obměny převážné většiny komponent NDK tak, abychom si udrželi krok s okolním světem, hrozí, že Národní digitální knihovna nebude dlouhodobě schopná poskytovat své služby čtenářům ani knihovníkům. Změny v legislativě EU

v oblasti ochrany osobních údajů, notoricky známé GDPR, se Národní digitální knihovny nepřímo také dotkne v souvislosti s možnou budoucí náhradou knihovnického systému Aleph, bez kterého jakékoli zpracovávání fondů Národní knihovny není možné.

Vedle uvedeného existují také „nové“ skutečnosti, které dávají podněty k implementaci nových funkcí a vlastností do systému Národní digitální knihovny. I když nejde o udržitelnost v striktním výkladu, je nutné se jimi také zabývat. Národní knihovna usiluje o zavedení povinnosti pro vydavatele e-born publikací předávat je tak, jako stále platí povinnost předávání povinného výtisku. S tím souvisí na straně Národní knihovny potřeba schopnosti takové soubory přijmout, zpracovat, ukládat a třeba i zpřístupnit.

Jedním se záměrů Národní digitální knihovny je skenování ve větším rozlišení než v současných 300 dpi. Protože tím ale skokově naroste datový objem v lince, musí se nejdříve vyřešit nákup nových úložných kapacit, možná i navýšení výpočetního výkonu a infrastruktury pro datové přenosy mezi místy zpracování. Chceme se také zaměřit na zvyšování kvality samotné produkce, například zvýšit úroveň kontroly barevného podání zavedením používání ICC profilů.

Vše výše uvedené má přirozeně finanční aspekt. Jak Studie proveditelnosti, tak pozdější upřesnění řešící udržitelnost, předpokládá, že přibližně po šesti letech provozu linky bude vhodné obnovit prakticky všechny HW prostředky. Vzhledem k tomu se neočekává významné zvyšování celkového nominálního výkonu linky, obnova nemá za cíl provést výkonový upgrade, ale jen zajistit patřičnou dostupnost hardwarových prostředků. Protože ale po celou dobu digitalizace vznikají nová a nová data nemalých objemů, bylo v rámci doby povinné udržitelnosti načasováno na rok 2019 zdvojnásobení úložné kapacity LTP. Ve finančních ukazatelích se předpokládá, že náklady na udržení NDK v provozu od 1. ledna 2015 do 31. prosince 2019 budou ve výši 260 600 000 Kč, prakticky 90% částky vynaložené na realizaci samotného projektu. Tato suma ovšem logicky nezahrnuje implementaci nových funkcí a změn softwarového vybavení, jejichž potřeba vznikla později. Na rozdíl od fáze realizace projektu, financované převážně z prostředků IOP, veškeré náklady na udržitelnost nelze z evropských fondů financovat. Prostředky si tedy musí Národní knihovna shánět sama ve svém rozpočtu, případně z dotací nebo jednorázových příspěvků zřizovatele knihovny, kterým je Ministerstvo kultury ČR.

A samozřejmě najít dodavatele. Národní knihovna ČR se svým zaměřením především ke knihovnictví, správě svých fondů a vědecké činnosti, nemá dostatečné kapacity pro

vývoj složitých a komplexních ICT systémů obdobného rozsahu, jako je Národní digitální knihovna provozovaná v Národní knihovně ČR. Řadu služeb, které si možná jiné instituce řeší vlastním vývojem nebo IT oddělením, tak musí NK ČR nakupovat zvenčí, to zase klade nároky na řídící a koordinační činnost. Přesto všichni v Národní knihovně věříme, že se nám podaří udržet systém Národní digitální knihovny nejen do prosince 2019, ale ještě pěknou řádku dalších let, aby mohla dále produkovat nové a nové přírůstky do digitální knihovny a splnit původní záměry, vládní strategii i společenskou poptávku.

Perspektivy digitální archivace v archivech mimo státní archivní síť na příkladu Univerzity Karlovy

Zdeněk Vašek, Petr Cajthaml, Eliška Pavlásková, Ústav dejín Univerzity Karlovy a Archív Univerzity Karlovy, Praha, ČR

Abstrakt

Příspěvek popisuje aktuální situaci a možnosti řešení problematiky ukládání digitálních archiválií v České republice. Věnuje se zejména potřebám archivů zřizovaných veřejnoprávními institucemi a stojících mimo státní archivní síť (dle české legislativy „specializovaných archivů“). V České republice je preferováno centralizované řešení v podobě jediného státního digitálního archivu, který je součástí Národního archivu v Praze. Zřizovatelům specializovaným archivům však archivní legislativa dává (v praxi dosud nevyužitou) možnost vybudovat vlastní digitální archivy. Specializované archivy mají totiž specifické požadavky na správu a zejména na přístup k uloženým digitálním archiváliím, kterým centralizované řešení nedokáže plně vyhovět. V příspěvku budou představeny legislativní a praktické požadavky na tyto archivy a návrh konkrétního řešení, které je připravováno v Archivu Univerzity Karlovy.

Abstract

The paper describes the current situation and possibilities of solving the issue of digital archival storage in the Czech Republic. The paper focuses mainly on needs of archives founded by public institutions and operating outside a network of state archives (according to legislation “Specialized archives”). In the Czech Republic is preferred a centralized solution of single digital archive which is a part of the National Archive in Prague. However, specialized archives are allowed according the legislation (in practice not yet used) to build their own digital archives. Specialized archives have specific requirements for the management of digital documents, and in particular the access to them, which the centralized solution can not fully accommo-

date. The paper presents the legislative and practical requirements for these archives and the proposal for a concrete solution prepared in the Archive of Charles University.

Úvod

Digitální dokumenty jsou již běžnou součástí života současné společnosti. Není proto nijak překvapující, že bylo nutné začít uvažovat, jak zajistit jejich dlouhodobé uchování poté, co se z nich dle platné legislativy staly archiválie. Úloha archivů ochránit kulturní dědictví a dokumenty o vývoji společnosti se rozšířila i na segment digitálních dokumentů. Vedle toho samozřejmě v mnohých archivech probíhá digitalizace za účelem ochrany a zpřístupnění klasických archiválií. Nicméně v tomto případě zůstávají digitalizáty pouze v pozici kopií na rozdíl od born-digital dokumentů, které je nutné chápat jako originály a zajistit jim v souladu s legislativou patřičnou ochranu.

Česká archivní síť se na potřebu správy digitálních archiválií připravuje dlouhodobě, de facto od počátku 21. století. Zásady dlouhodobého uchovávání jsou dostatečně známé a prověřené, ale jejich realizace v jednotlivých typech institucí se liší podle jejich potřeb. To platí i pro archivy, jejichž nároky jsou dány mj. jejich různými typy.

Archivace digitálních dokumentů je v českém prostředí již realitou, tyto dokumenty jsou součástí archivních fondů. Archivy aktuálně disponují nástroji na dlouhodobou ochranu předaných dokumentů anebo tyto nástroje zavádějí. Při jejich vývoji se samozřejmě musí zamýšlet nad jejich koncepcí a formulovat ji tak, aby zajistili jejich udržitelnost. Potřeba dobře plánované udržitelnosti není diktována jen logickým a ekonomickým hlediskem, ale vychází ze samotné podstaty archivů a je dána jedinečností uložených dokumentů. Příspěvek si klade za cíl představit potřeby archivů stojících mimo síť státních archivů a jejich specifické potřeby ohledně archivace digitálních dokumentů. Na příkladu Univerzity Karlovy (dále UK) pak popíše zvolené řešení těchto potřeb na příkladu plánovaného informačního systému.

Digitální dokumenty v ČR jako archiválie

Česká archivní legislativa (především zákon č. 499/2004 Sb. o archivnictví a spisové službě v platném znění) pracuje s pojmem „archiválie v digitální podobě“, pod tímto

pojmem jsou chápány dokumenty vzniklé jako originál v digitální podobě (born-digital dokumenty), které byly vybrány za archiválie a jsou uloženy a spravovány v některém z archivů zřízených dle výše uvedeného zákona (viz § 15, odst. 3. zákona č. 499/2004 Sb.). „Archiváliemi v digitální podobě" nejsou digitalizáty papírových (analogových) archiválií, jejich ukládání a péče o ně není v české archivní legislativě jako celek řešena.

Na nutnost archivace dokumentů v digitální podobě začali čeští archiváři narážet již v polovině 90. let, kdy se stále častěji setkávali s dokumenty trvalé historické hodnoty, které existovaly jen v digitální podobě. Na počátku tohoto století byly za účasti Odboru archivní správy MV ČR, Českého vysokého učení technického a Národního archivu (dále NA ČR) realizovány první výzkumné projekty v oblasti možného technického a procesního řešení digitální archivace. Důležitým milníkem bylo vládní usnesení č. 11/2004 k dlouhodobému uchovávání a zpřístupňování dokumentů v digitální podobě, které uložilo ministerstvu vnitra vytvořit při NA ČR odborný tým, který zajistí vybudování celostátního digitálního archivu (Usnesení vlády ČR č. 11/2004). Původní harmonogram byl velmi optimistický, počítal se zahájením rutinního provozu digitálního archivu od začátku roku 2008.

Přestože v uplynulých patnácti letech byly diskutovány varianty vybudování Národního digitálního archivu (dále NDA) jako de facto samostatné instituce nezávislé na stávajících archivech pro ukládání „papírových" archiválií a vydání samostatného zákona o digitální archivaci, nedošlo nakonec k jeho vydělení. Problematika ukládání digitálních archiválií byla legislativně řešena v několika novelách archivního zákona a NDA, který po několika neúspěšných projektech zahájil zatím omezený provoz před třemi lety, je organizační součástí oddělení NA ČR pověřeného péčí o nejnovější archiválie a dozorem nad výkonem spisové služby u centrálních úřadů České republiky.

Ačkoli česká archivní legislativa umožňuje vznik dalších digitálních archivů, zůstává NDA v roce 2018 jediným digitálním archivem v české archivní síti a zajišťuje uložení digitálních archiválií pro všechny české archivy. NDA funguje jako digitální repozitář pro archiválie ostatních archivů. NDA odpovídá za jejich uložení a dlouhodobou ochranu, samotná správa (obnášející výběr, evidenci, zpracování a v budoucnu i zpřístupnění) digitálních archiválií nadále zůstává v kompetenci jednotlivých archivů podle jejich věcné a územní příslušnosti.

Pro archiváře ostatních archivů je klíčovým nástrojem NDA Národní archivní portál, který poskytuje soubor nástrojů pro výběr dokumentů, které se mají stát archiváliemi, validaci metadat úředních dokumentů, případně tvorbu metadat neúředních dokumen-

tů a uložení vybraných dokumentů jako digitálních archiválií v NDA. Vcelku rutinně jsou realizovány výběry a ukládání úředních digitálních dokumentů a digitálních metadat papírových dokumentů, které jsou evidovány v elektronických systémech spisových služeb. Zatím není možné uložené digitální archiválie archivně zpracovávat a velmi omezená je možnost jejich zpřístupnění (ať už pro archivy, do jejichž péče přísluší, či pro původce).

Nestátní archivy a digitální archiválie

Archivnictví v České republice jako celek spadá do gesce Ministerstva vnitra, které jej řídí prostřednictvím svého odboru archivní správy. Základem archivní sítě je dvoustupňová soustava státních archivů tvořená NA ČR a státními oblastními archivy, jejichž organizační součástí jsou státní okresní archivy. Tyto archivy se starají o archiválie všech původců, kteří si nezřídili vlastní archivy. V oblasti ukládání digitálních archiválií státní archivy musí dle platného zákona ukládat v NDA. Přestože jsou digitální archiválie státních archivů uloženy v jediném existujícím digitálním archivu, jejich správa (v současné době především jejich výběr a evidence) je nadále zajištěna archivy podle jejich územní a věcné příslušnosti. Jednotlivé archivy přitom využívají nástrojů a workflow, které jsou součástí Národního archivního portálu, který je spravován NA ČR.

Státní archivní síť doplňují další typy archivů (specializované, bezpečnostní, soukromé a územních správních celků), které slouží primárně svým zřizovatelům. Ministerstvo vnitra je přímo neřídí, vykonává nad nimi jen metodický dohled.

Zpravidla nejde o samostatné instituce, ale o organizační součásti svých zřizovatelů. Ti musí pro zřízení svých archivů splnit řadu personálních, technických a organizačních požadavků, jejich splnění je prokazováno před ministerstvem vnitra v rámci akreditačního řízení.

V oblasti ukládání digitálních archiválií mají nestátní archivy více možností než státní, které musí povinně ukládat v NDA.

Pro nestátní archivy je to jen jedna ze tří možností. Druhou a pro náš příspěvek nejdůležitější možností je možnost vybudovat v rámci nestátních archivů vlastní digitální archiv, který zajistí (pro zřizovatele nestátního archivu) většinu funkcionalit NDA. V archivní legislativě je tato varianta popsána jako právo zřizovatelů nestátních archivů získat oprávnění ukládat archiválie v digitální podobě. Zákon toto právo dává všem

zřizovatelům již akreditovaných archivů, pokud zřizovatel svůj akreditovaný (tj. archiv určený pro ukládání "papírových archiválií") archiv ještě nezřídil, musí zároveň požádat o akreditaci "papírového" archivu. Tím je mj. zabezpečena jednotná správa papírových a digitálních archiválií zřizovatele a jeho dostatečné kompetence v oblasti odborné péče o archiválie.

Legislativní požadavky na nestátní digitální archiv jsou spíše obecného rázu, podrobněji jsou popsány jen technické požadavky na umístění technologií digitálního archivu (mj. platí požadavek na minimální, padesáti kilometrovou, vzdálenost hlavního a záložního úložiště), součástí řízení o oprávnění však není žádný audit podle mezinárodně uznávaných norem. Požadavky na získání oprávnění pro ukládání archiválií v digitální podobě jsou shrnuty v § 60a zákona č. 499/2004 Sb. Jedinou prováděcí normou je v současné době Vzorový provozní řád digitálního archivu (Věstník ministra vnitra, částka 65/2012), který poměrně detailně upravuje procesní náležitosti digitálního archivu a v obsahové rovině předepisuje náležitosti archivního informačního balíčku.

Samotné oprávnění pro ukládání digitálních archiválií uděluje Ministerstvo vnitra, důležitá role v této oblasti náleží Národnímu archivu jako zřizovateli NDA. Jedním z mála konkrétních požadavků na strukturu ukládaných dat je požadavek na kompatibilitu s NDA, vyjádřená požadavkem na úspěšný přenos zkušební dávky minimálně 50 archivních informačních balíčků do NDA. Součástí žádosti o získání oprávnění pro ukládání digitálních archiválií dále musí být podklady, na jejichž základě žadatel získal akreditaci svého běžného archivu, popis způsobu uložení archiválií v digitální podobě, koncepce dlouhodobé ochrany a uchovávání, identifikace ukládaných archiválií v digitální podobě, seznam archivních metadat a popis evidence archivních souborů a původců a návrh provozního řádu digitálního archivu.

Z platné legislativy je zřejmé, že nestátní digitální archiv bude muset obsahovat řadu funkcionalit a rolí nad rámec funkčnosti klasického LTP archivu (nad rámec ukládání a dlouhodobé ochrany digitálních dokumentů). Zřizovatel digitálního archivu bude mimo jiné muset disponovat vlastním archivním portálem, který zajistí většinu funkcionalit NDA, především bude disponovat nástroji pro výběr dokumentů za archiválie, nástroji pro zpracování uložených archiválií ("archivní katalogizaci") a zpřístupnění uložených archiválií veřejnosti.

Dosud však k vybudování žádného „nestátního" digitální archivu nedošlo. Žádost o získání oprávnění pro ukládání archiválií v digitální podobě podala jen jediná instituce – Masarykova univerzita v Brně jako zřizovatel Archivu Masarykovy univerzity. Rozhodování o této žádosti dosud nebylo ukončeno (září 2018). Praktické kroky k vy-

budování vlastního digitálního archivu podniká hlavní město Praha jako zřizovatel Archivu hlavního města Prahy, Škoda Mladá Boleslav, jako zřizovatel akreditovaného soukromého archivu a v delším časovém horizontu o něm uvažují České vysoké učení technické v Praze a Univerzita Palackého v Olomouci.

Zřizovatel nestátního archivu v České republice má tedy v oblasti ukládání digitálních archiválií možnost volby mezi ukládáním v NDA nebo zřízením vlastního digitálního archivu. Legislativa připouští i třetí, zatím jen teoretickou, možnost, kdy by zřizovatel nestátního archivu ukládal své digitální archiválie na základě smlouvy v digitálním archivu, který je součástí jiného nestátního archivu. Tato možnost je však zatím prakticky nerealizovatelná a proto se jí dále nebudeme věnovat.

Archivace digitálních dokumentů v prostředí vysokých škol

Veřejné vysoké školy v České republice v minulých desetiletích úspěšně digitalizovaly většinu svých úředních a správních agend a díky tomu zpracovávají velké množství digitálních dokumentů, řada z nich má trvalou historickou hodnotu a je nutné zajistit jejich budoucí archivaci. Zhruba třetina českých vysokých škol zřídila své specializované archivy, a proto se bude v budoucnu aktivně podílet na správě svých digitálních archiválií, pro ostatní vysoké školy archivaci zajišťují státní oblastní archivy.

Důležitým zdrojem zkušeností v oblasti digitální archivace v prostředí vysokých škol byl centralizovaný rozvojový projekt Ministerstva mládeže, školství a tělovýchovy probíhající v letech 2015 až 2017, jehož tématem bylo prohloubení správy a zajištění budoucí digitální archivace dokumentů veřejných vysokých škol. Na projektu se podílela většina českých veřejných vysokých škol (hlavním řešitelem byla Masarykova univerzita). Důležitou částí projektu však bylo vyjasnění podmínek pro zřízení vlastního digitálního archivu, vysoké školy v této věci komunikovaly s Ministerstvem vnitra a NA ČR. V rámci projektu vznikl metodický materiál řešící možnosti digitální archivace v oblasti předarchivní péče i vzniku samotných digitálních archivů (Pichl et al., 2015). Jedním z výstupů byla metodika pro vyřazování a archivaci úředních dokumentů ve skartačním řízení (Cajthaml, 2017).

Masarykova univerzita v rámci projektu zpracovala koncepci a osnovu technického řešení svého budoucího digitálního archivu.

Digitální dokumenty Univerzity Karlovy a Archiv UK

Univerzita Karlova (UK) je největší univerzitou v České republice a jako taková je producentem velkého množství digital-born dokumentů různých typů, které jsou spravovány v řadě informačních systémů a digitálních repozitářů. Z hlediska digital-born dokumentů lze identifikovat dva hlavní okruhy systémů, ve kterých vznikají a jsou spravovány.

Prvním jsou specializované ("agendové") informační systémy, které jsou v prostředí UK postupně zaváděny a které v průběhu posledních zhruba dvaceti let nahrazují dříve papírové agendy. Druhou zásadní oblastí je klasická evidence a správa úředních dokumentů, tedy spisová služba, kterou univerzita vykonává povinně v elektronické podobě v elektronickém systému spisové služby. V tomto systému jsou evidovány a spravovány jak digitální, tak analogové (papírové) dokumenty a spisy. Systém ukládá digitální dokumenty a jejich metadata, v případě analogových a digitálních dokumentů jen jejich metadata.

Z hlediska dlouhodobého uchovávání je zásadním přínosem elektronické spisové služby skutečnost, že jejími výstupy jsou SIP dle Národního standardu pro elektronické spisové služby (NSESS). NSESS specifikuje podobu metadatového záznamu dokumentu nebo spisu a je založen na standardu METS. Ve spisové službě mohou být evidovány digitální, analogové i hybridní (obsahující analogové i digitální části) dokumenty. Výstupní datové formáty dokumentů v digitální podobě jsou určeny §23 vyhlášky č. 259/2012 Sb. o podrobnostech výkonu spisové služby, která specifikuje formáty dat pro textové, obrazové a audiovizuální dokumenty a pro databáze. Formáty v obecné rovině odpovídají mezinárodně přijímaným požadavkům na formáty vhodné k dlouhodobé archivaci digitálních objektů (Sustainability Factors, 2018). V případě analogových dokumentů spravovaných v elektronické spisové službě výstupní SIP obsahuje jen metadata těchto dokumentů. Vyřazování a výběr dokumentů k archivaci ze spisové služby po konci jejich životního cyklu je řešen v rámci standardizovaného workflow – skartační řízení. Archiváři na základě návrhu vybírají dokumenty určené k archivaci a povolují ničení těch, které trvalou hodnotu nemají.

Velké množství digitálních dokumentů (a informací, které nemají formu klasických dokumentů) je spravováno také ve výše zmíněných specializovaných informačních systémech. V tuto chvíli již do prakticky všech procesů, které na univerzitě probíhají, více či méně zasahují digitální technologie. Záznamy o studentech i o studiu samot-

ném jsou uchovávány ve studijním informačním systému, samostatné systémy existují pro administraci grantů a v neposlední řadě UK disponuje i repozitářem závěrečných prací, kde jsou shromažďovány studentské práce v digitální podobě (Pavlásková, 2017). Tyto systémy však nejsou zpravidla standardizovány a neumožňují předávání spravovaných dokumentů (respektive spravovaných informací) k uložení v nezávislém digitálním archivu,

Cílovým řešením proto bude integrace těchto systémů s elektronickým systémem spisové služby, tak aby zpracovávané dokumenty byly evidovány spisovou službou. K jejich vyřazování a archivaci pak bude docházet prostřednictvím výše popsaného workflow skartačního řízení. V případech, kdy integrace není možná, se Archiv UK snaží ovlivnit výslednou podobu dat tak, aby byla produkována ve formátech vhodných k dlouhodobé archivaci a byla opatřena alespoň základními metadaty, která umožní další zpracování a dlouhodobé uložení (Cajthaml, Pavlásková (2018).

Archiv UK a Archivní informační systém

Za výkon archivní péče je v rámci UK odpovědný Archiv UK, který je součástí Ústavu dějin a Archivu UK, který je akreditovaným specializovaným archivem, jehož historie sahá až do 14. století. Neplní pouze funkci archivu, je i vědeckou institucí zabývající se dějinami Univerzity Karlovy a dějinami školství a vzdělanosti v českých zemích. Jeho fondy nyní zahrnují především papírové dokumenty, které jako celek představují 8 km archiválií. V současné době má 18 zaměstnanců, z nichž tři se věnují přípravě možností vlastní digitální archivace, na jejíž technické stránce úzce spolupracují i s Ústavem výpočetní techniky UK (ÚVT). Důležitou součástí práce Archivu UK předarchivní péče, která obnáší zejména dohled nad správou dokumentů na Karlově univerzitě a jejich výběr k trvalému uložení (archivaci). V oblasti digitálních dokumentů se Archiváři UK podíleli na výběru a zavádění elektronického systému spisové služby na UK, standardizaci dalších významných informačních systémů (např. správy elektronických kvalifikačních prací).

Archiv UK v současné době uchovává omezené množství digitálních archiválií (v objemu desítek GB), zpravidla neúřední provenience, které jsou ukládány provizorně bez odpovídající správy. Do doby vybudování případného vlastního digitálního archivu UK budou digitální archiválie ukládány v NDA, první skartační řízení, jehož výsledkem bude uložení digitálních archiválií v NDA, proběhne na konci roku 2018.

Velkou zkušeností a výchozím bodem pro problematiku LTP byl pro Archiv UK stále ještě probíhající projekt Studenti pražských univerzit 1882-1945, kde již bylo digitalizováno více než 123 000 stran (Cajthaml, Vašková, 2017). Digitalizáty jsou ukládány a spravovány v diskovém úložišti, v úložišti CESNET a na fyzických nosičích. Zpřístupňovány jsou v rozhraní vyvinutém přímo na míru projekt pracovníky ÚVT.

Archivu UK vykonává pro svého zřizovatele řadu správních funkcí. Pro zřizovatele je proto důležité, aby archiv zajistil nejen dlouhodobé uchovávání dokumentů, ale také rychlé a komfortní vyhledávání v předaných dokumentech, a to i v nezpracovaných archiváliích (tedy v archiváliích, které neprošly procesem archivní katalogizace, jejímž výsledkem je archivní pomůcka – systematizovaný popis archivního fondu; zpracování archiválií představuje obvykle dlouhodobý proces trvající řádově i několik let) a následný rychlý přístup k vyhledaným dokumentům. Uložení digitálních archiválií pouze v NDA, který poskytuje jen omezený přístup k předaným dokumentům, tak pro UK nepředstavuje funkční řešení. Vedle toho je třeba zdůraznit relativně velký objem různorodých digitálních archiválií popsaných výše a vznikajících z činnosti UK.

Archiv UK stál před otázkou, jakým způsobem dostát legislativním nárokům na uložení digitálních archiválií a současně zabezpečit alespoň minimální požadavky UK na využívání archiválií v další činnosti. Jediným vhodným a prakticky realizovatelným řešením se na základě podrobné studie ukázala nutnost vybudování vlastního nástroje, který by odpovídal specifickým požadavkům univerzity, zejména v oblasti přístupu k uloženým dokumentům. Nástroj by v budoucnu (po nutném rozšíření a splnění podmínek) měl umožnit i získání oprávnění k uložení archiválií v digitální podobě. Přidanou hodnotou je samozřejmě i možnost zpřístupnění archiválií badatelům přímo v systému přizpůsobeném požadavkům uživatelů z řad akademické i laické veřejnosti. Nezbytnou součástí je i řízení přístupu k dokumentům podle typu uživatelských práv. To je zvláště důležité v situaci, kdy část uživatelů tvoří interní pracovníci zřizovatele a část badatelská veřejnost.

Jakékoliv nástroje pro ukládání a správu digitálních dokumentů, ať už s nástroji LTP nebo bez nich, představují komplexní systémy náročné na hardwarové a softwarové součásti i odborné zázemí provozovatele. Vedle nároků ale přináší též mnohé výhody. V případě Archivu UK jimi jsou zejména:

- možnost zpracování specifických typů archiválií, které vychází z univerzitní praxe a jsou typické jen pro tento typ původců,
- rychlé zpřístupnění schopné v dostatečné míře reagovat na potřeby vyplývající z běžné agendy UK,

- plná podpora správních činností UK a umožnění vyhledávání nezpracovaných archiválií,
- kontrola nad způsobem uložení a správy archiválií produkovaných UK,
- zachování jednoho institucionálního zázemí pro oblast ukládání a správy digitálních archiválií, možnost nezávislosti na ne zcela vyhovujících nástrojích NDA,
- jednotný informační nástroj pro výběr, příjem, správu a zpřístupnění archiválií bez ohledu na to zda mají digitální či analogovou formu.

Z výše popsaných důvodů se UK v roce 2017 rozhodla získat softwarový nástroj určený ke komplexní podpoře činností Archivu UK, jehož součástí bude též ukládání digitálních dokumentů. Jeho pracovní název zní Archivní informační systém (dále AIS). V rámci technických, personálních i finančních možností se UK rozhodla vybudovat nástroj, který bude splňovat její praktické potřeby, avšak nenaplní všechny formální předpoklady pro samostatný digitální archiv. V současnosti (září 2018) je před dokončením zadávací dokumentace pro výběr dodavatele tohoto systému. Provoz systému na UK se předpokládá od roku 2020. V první fázi provozu systému nebude UK usilovat o získání oprávnění k ukládání digitálních archiválií ve smyslu archivní legislativy. Rozhodnutí o dalším rozšíření a případné žádosti o oprávnění učiní na základě získaných zkušeností. Od legislativních požadavků na digitální archiv se UK odchýlí především v otázce uložení. Nebude budovat geograficky oddělené úložiště a spokojí se pouze s jedním úložištěm se dvěma kopiemi dat (z toho jedna offline). LTP ochranu zajistí prostřednictvím NDA, do kterého bude v souladu s legislativou odesílat informační balíčky s archiváliemi.

Cílem UK je získat nástroj, který umožní výběr, příjem, evidenci, zpracování, uložení a zpřístupnění digitálních a analogových archiválií (včetně jejich případných digitálních kopií) a informační podporu činností archivu. Nechce se omezit jen prosté zajištění uložení, které by nezajistilo další nutné činnosti archivu a znamenalo by v budoucnosti potřebu dalších informačních systémů a jejich náročnou integraci.

Důvodů snahy o takto komplexní nástroj je hned několik. Archiv UK bude v budoucnu muset spravovat nejen digitální archiválie, ale stále součástí jeho činností zůstanou i klasické papírové archiválie. Jejich správu nelze oddělit, není proto optimální, aby vedle sebe existovaly dva systémy. AIS proto musí podporovat i správu záznamů o analogových archiváliích. Navíc je nutné připomenout existenci tzv. hybridních archiválií, jejichž součástí je digitální i analogový dokument (případně analogový dokument a digitální původcovská metadata) a tudíž z podstaty věci zpracování v odlišných systémech vylučují. Z ekonomického pohledu by bylo značně nevýhodné provozovat

dva systémy, byť by mohly být podobné. Jejich integrace do společného rozhraní by představovala trvalé ekonomické a technologické nároky, které s jednotným nástrojem odpadají. Vedle vlastní správy musí archiv disponovat nástrojem pro zpracování archiválií. I tento nástroj by měl být integrován do AIS. V opačném případě by se jednalo o další systém, jehož integrace by musela být řešena samostatně. Správa rozsáhlých souborů analogových archiválií je bez moderních nástrojů také značně komplikovaná. Je tedy logické integrovat tyto činnosti do jednoho nástroje, který navíc pokryje i oblast vyhledávání, zpřístupnění a ukládání.

Nároky na oficiálně akreditovaný digitální archiv jsou dané legislativou. V odborné komunitě je jako nejuznávanější soubor principů pro budování systému pak dlouhodobého uchovávání chápána norma ČSN ISO 14721. Zákon i norma jsou poměrně striktní a požadují řadu kroků. UK v první fázi realizace vlastního řešení definovala škálu nároků, které odpovídají situaci UK a dobré praxi.

Systém by měl fungovat jako webová aplikace, která umožní informační podporu všech činností archivu. V jednom prostředí bude možné administrovat celý životní cyklus archiválie, který musí odpovídat procesům realizovaným v celé archivní síti, ve které je však rozdělen do více nástrojů.

Základní vrstvou, nad kterou bude systém vybudován, je software úložiště. To bude v rámci systému fungovat víceméně nezávisle, do zbytku systému bude odesílat datové soubory a zpět je přijímat. Dokumenty v rámci úložiště budou organizovány do archivních informačních balíčků. Úložiště zajistí bezpečné uložení digitálních archiválií a dalších dokumentů na bitstreamové úrovni.

Nad tímto úložištěm by měl běžet zbytek nástroje, který bude rozdělen do tří celků a několika administrativních a podpůrných nástrojů sloužících k řízení a nastavování procesů a realizaci kontrolních procesů, validací apod. Tři celky byly pracovně označeny jako Výběr, Správa a Zpřístupnění. V rámci těchto celků budou zapojeny nástroje a moduly umožňující realizaci požadovaných operací popsaných níže. Informační systém by měl fungovat na modulární bázi, aby v budoucnu umožňoval nahrazení jednotlivých zapojených nástrojů.

V rámci celku Výběr bude umožněna realizace výběru dokumentů v rámci procesů skartačního a mimoskartačního řízení. V něm budou úřední i neúřední dokumenty zkontrolovány, zpracovány, případně bude doplněn jejich popis a poslány do celku Správa a zároveň uloženy na úložiště (v případě analogových archiválií budou v úložišti uložena jen metadata, Archiv UK bude nadále přejímat archiválie v listinné podo-

bě). Vedle archivních informačních balíčků bude v celku Správa existovat databáze se záznamy o archiváliích, pomůckách a událostech, které v archivu proběhly. Budou zde zapojeny nástroje pro operace s archiváliemi podle archivní legislativy a nástroje pro vyhledávání a zpracování archiválií. V celku Zpřístupnění bude zajištěn přístup k uloženým archiváliím a také podpora komunikace s badateli, vedení jejich účtů a další náležitosti. Součástí systému bude i webové rozhraní (portál) pro komunikaci s uživateli. Celky by měly být navzájem odděleny, komunikace bude realizována jen vybranými rozhraními, celek Správa bude do značné míry izolován od vnějšího světa a přístup bude přísně kontrolován. Důvodem je především bezpečnost uložených archiválií a údajů v nich obsažených, zejména pokud jde o osobní údaje. Konkrétní realizace funkčních požadavků bude předmětem vývoje, stejně jako otázka licencí k informačnímu systému atd.

Závěr

Takto pojatý systém nemá v rámci ČR ve své komplexitě srovnání, jeho koncepce však vychází z logických požadavků na integraci činností, které jsou dnes prováděny různými nástroji. AIS by měl zajistit všechny potřeby, které na archivy klade legislativa ČR v oblasti zpracování, správy a zpřístupnění archiválií (analogových i digitálních) a navíc fakticky zajistit bezpečné uložení digitálních archiválií. Propojením se správou analogových archiválií umožní zjednodušení celé agendy a tím i snížení náročnosti provozu a omezení rizik, které hrozí z důvodu přechodů mezi různými systémy. Zjednodušení a integrace nástrojů do jednotného informačního systému je z pohledu UK důležitým krokem k zajištění budoucnosti archivace digitálních dokumentů a její udržitelnost do budoucna.

Zdroje:

CAJTHAML, Petr a Lenka VAŠKOVÁ. Studenti pražských univerzit 1882-1945: Digitalizace v Archivu Univerzity Karlovy. In: *Archivum Trebonese*. Třeboň: Státní oblastní archiv v Třeboni, 2017, s. 204-215. ISBN 978-80-906860-1-4.

CAJTHAML, Petr a Eliška PAVLÁSKOVÁ. Evolution of Electronic Management Systems, Digital Archiving and Czech Universities: From Student Information Systems to Digital Records Management and Long Term Preservation. In: BLECHER, Jens, Sabine HAPP a Juliane MIKOLETZKY. *Normen und Ethos: Schreiben Archiva-*

rinnen und Archivare Geschichte?. Leipzig: Leipziger Universiätsverlag, 2018, s. 193-202. ISBN 978-3-96023-188-2.

CAJTHAML, Petr. *Metodika pro skartaci dokumentů evidovaných v systémech elektronické spisové služby v prostředí vysokých škol* [online]. Praha: Ústav dějin a archiv Univerzity Karlovy, 2017 [cit. 2018-09-30]. Dostupné z: https://is.muni.cz/digitalniuniverzitaMetodika_skartace_UK.pdf

ČSN ISO 14721. Systémy pro přenos dat a informací z kosmického prostoru – Otevřený archivační informační systém – Referenční model. 1. vyd. Praha: Úřad pro technickou normalizaci, metrologii a státní zkušebnictví, 2014.

ČSN ISO 16363. Systémy pro přenos dat a informací z kosmického prostoru – Audit a certifikace důvěryhodných digitálních úložišť. 1. vyd. Praha: Úřad pro technickou normalizaci, metrologii a státní zkušebnictví, 2014.

HUTAŘ, Jan, Andrea MIRANDA, Eliška PAVLÁSKOVÁ, Zdeněk VAŠEK a Zdeněk HRUŠKA. *Metodika logické ochrany digitálních dat* [online]. Praha: Knihovna Akademie věd, 2017 [cit. 2018-09-29]. Dostupné z: http://hdl.handle.net/11104/0282107

Národní standard pro elektronické systémy spisové služby; Věstník Ministerstva vnitra, částka 57/2017. Dostupné z: http://www.mvcr.cz/soubor/vestnik-mv-57-2017-oznameni-ministerstva-vnitra-kterym-se-zverejnuje-narodni-standard-pro-elektronicke-systemy-spisove-sluzby.aspx

PAVLÁSKOVÁ, Eliška. From the Dissemination of Electronic Theses and Dissertations to Their Long-term Archiving. In: *10th Conference on Grey Literature and Repositories: proceedings* [online]. Prague: National Library of Technology, 2017, s. 33-44 [cit. 2018-09-27]. ISSN 2336-5021. Dostupné z: http://nrgl.techlib.cz/conference/conference-proceedings

PICHL, Marek, Miroslav KŘIPAČ, Jitka BRANDEJSOVÁ, Růžena ZEMANOVÁ a Michal BRANDEJS. *Metodika dlouhodobého ukládání a archivace digitálních dokumentů* [online]. Brno: Fakulta informatiky Masarykovy univerzity, 2015 [cit. 2018-09-29]. ISBN 978-80-210-8113-0. Dostupné z: https://is.muni.cz/repo/1322181/Metodika_dlouhodobeho_ukladani_a_archivace_digitalnich_dokumentu.pdf

Sustainability Factors. Sustainability of Digital Formats: Planning for Library of Congress Collections [online]. Washington: Library of Congress, 2018 [cit. 2018-09-

27]. Dostupné z: https://www.loc.gov/preservation/digital/formats/sustain/sustain.shtml.

Usnesení vlády České republiky ze dne 7. ledna 2004 č. 11 k dlouhodobému uchovávání a zpřístupňování dokumentů v digitální podobě. [cit. 2018-09-27] Dostupné též z: https://www.nacr.cz/wp-content/uploads/2015/12/chimera_usneseni.pdf.

Vyhláška č. 259/2012 Sb., o podrobnostech výkonu spisové služby v platném znění. [cit. 2018-09-27] Dostupné též z: https://www.zakonyprolidi.cz/cs/2012-259

Vyhláška č. 645/2004 Sb., *kterou se provádějí některá ustanovení zákona o archivnictví a spisové službě a o změně některých zákonů* v platném znění. [cit. 2018-09-27] Dostupné též z: https://www.zakonyprolidi.cz/cs/2004-645

Vzorový provozní řád archivu oprávněného k ukládání archiválií v digitální podobě (pouze vybraná ustanovení, Věstník Ministerstva vnitra, částka 65/2012. Dostupné z: http://www.mvcr.cz/soubor/65-vmv-pdf.aspx

WANNER, Michal et al. *Základní pravidla pro zpracování archiválií* [online]. Druhé, opravené a rozšířené vydání. Praha: Odbor archivní správy a spisové služby MV, 2015 [cit. 2018-09-30]. ISBN 978-80-86466-78-1. Dostupné z: http://www.mvcr.cz/soubor/zakladni-pravidla-pro-zpracovani-archivalii-2015-cervene-vyznacenymi-zmenami.aspx

Zákon č. 499/2004 Sb., o archivnictví a spisové službě v platném znění. [cit. 2018-09-27]. Dostupné též z: https://www.zakonyprolidi.cz/cs/2004-499

Implementation of New Technologies to Ensure the Sustainability of Digital Content

Zoltán Lux, National Archives of Hungary, Budapest, Hungary

Abstract
Sustainability of Long-term Preservation has several aspects that can be discussed. Technology, management, policies, cost, etc. are all important factors to be considered. This article focuses on the technological aspect of sustainability of digital information which is becoming more and more complex, and appears in a rapidly growing variety of digital formats.

Relational databases, Data Warehouses (DW), OLAP (Online Analytical Processing) objects are produced in an increasing number and size which have to be preserved and kept understandable for long time. The National Archives of Hungary piloted in the frame of an EU founded international project how data warehouse concept can be applied in archiving and presenting archived relational databases.

The article will first review the already-used methods for archiving databases. Then it will be explained why the idea of using a Data Warehouse concept was raised, and how and in which parts of the archiving process can it be applied.

Introduction

Memory organizations have been thinking for quite a long time about, that they should something to do with the continually growing amount of digital content in order to be able to use it over the long term. Without ensuring the long-term accessibility and interpretability of different types of digital content, they would just be destroyed as if we just deleted them. That is why the OAIS reference model has been evolved and devel-

oped to ISO standard (ISO-14721) [1]. It already became a reference system for the development and maintenance of long-term digital archiving systems in archives. OAIS model identified specific functions (functional entities) which have to be provided for long-term digital archives and handles the digital content to be archived is handled as an Information Package (IP).

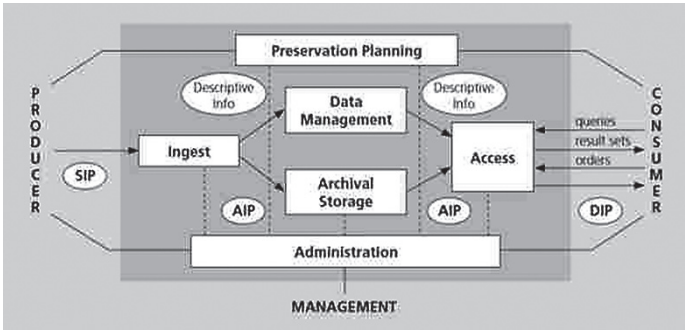


Figure 1: The Open Archival Information System (OAIS) Reference Model

Archiving Relational Databases

Public organizations and companies support their internal processes and collect data for their professional tasks in applications based on relational databases for many years. They also have started communicating with their clients via Internet based applications where the interactions are represented (documented) as data stored in relational databases. There are a big variety of such application software and RDBMSs behind them. It is impossible to maintain all these information systems for long term, so it was a task to work out, how the information, stored in those systems, could be archived for later use.

Databases are one of the main technologies that support information assets of organizations. They are designed to store, organize and explore digital information, becoming such a fundamental part of information systems that most would not be able to function without them. [2] The Swiss Federal Archives developed (2007) the Software Independent Archiving of Relational Databases (SIARD) format for archiving RDBMS on four internationally recognised standards: XML, SQL:2008, UNICODE and ZIP64, and it also became as Swiss E-Government Standard for archiving relational databases (eCH-0165). [3]

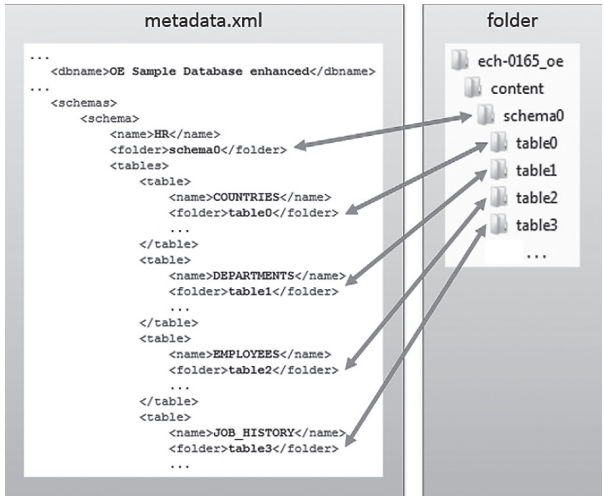


Figure 2: The Structure of the SIARD File. [4]

The SIARD format, basically, consists of a ZIP file that contains a hierarchy of folders and files of XML and XSD (XML Schema) format and the XML files inside the SIARD file hold database metadata information and contents.

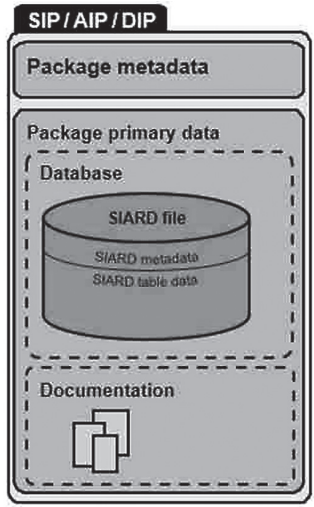


Figure 3: The Structure of an Information Package Containing a SIARD File. [5]

The Swiss Federal Archive developed a tool, called SIARD Suite, by which from specific RDBMS systems (DB2, Oracle, Microsoft SQL Server, MySQL, Microsoft Access), the contained data can be exported into a SIARD package. For finalizing the archival process the generated SIARD package will be packaged into an Archival Information Package (AIP) which will be contain additional metadata, optionally documentations.

Accessing Archived Relational Databases

SIARD suite is applicable for rendering and reusing archived RDBMS as well. However, only the individual tables can be browsed with it. User has to know very well the structure of the databases in order to retrieve the searched information. In the case, when the first relational database was submitted to the National Archives of Hungary, it was decided to develop a specific rendering tool to make the content available in a user friendly way. The archived SIARD package was imported into a live Oracle database and a simply Oracle APEX application was developed, which gave the user the ability to generate the most common queries and also creating new reports. Oracle APEX is for free, even if you use the only the Oracle Express version, but it needs not too much development effort.

It emerged at that point the Data Warehouse concept would be worth to try out both in rendering and in archiving phase of RDBMSs.

Transforming the structure of an operational RDBMS into a data warehouse can provide an environment in which an archived database can be presented to users, enabling them to run new queries which have never been thought of when the database was still in use. Usually this means that the data model will be de-normalized and so the structure will also be much more transparent for the user.

The implication of the DW concept to the archiving phase is that that the transformation of the original, usually normalized database, provides a simpler structure that is easier to understand. Of course it is a serious question whether the original database should also be stored, in all case, as a different manifestation within the archival information package or not.

Emerging Problems and Needs

Data warehouse is not only a tool by which RDMSs can be provided to user for further reuse. They are also produced in a great numbers at the organizations and in many cases they also contain historically valuable data, which could be difficult to re-produce because they were collected from different systems.

Data warehousing offers further additional functionalities and added values to the end user by supporting advanced use cases for OLAP and Data Mining.

During the E-ARK project (2014-2017) [6] the National Archives of Hungary piloted the applicability of the DW concept in archiving and presenting archived RDBMS to the users. [7]

Thinking further the DW concept, it was also piloted, how Business Intelligence tools can be used for analyzing data in archival context. To do this, the data model had to be transformed into multidimensional model and then into business model. The result was quite powerful, but it needed very specific tools. Furthermore, the way they (multidimensional and business model) are physically implemented in the different RDBMS, is particularly vendor specific and there is no widespread and accepted standard for this. [8]

Within the frame of the E-ARK project the SIARD 1.0 format has been further developed in SIARD Version 2.0. [9] Since the SIARD Suite has not yet been able to handle the new format [10], an existing other tool for a similar purpose has been further developed for exporting specified RDBMS [11] into SIARD 2.0 preservation format and importing such an export file back, into a live RDBMS. This tool was the Database Preservation Toolkit [12], developed by Keeps Solutions, and which was incorporated into the Hungarian archiving process.

Based on these experiences, the National Archives of Hungary also contributed to the further development of the Dissemination Information Package (DIP) specification [13] of the E-ARK project, in which all above mentioned problems are discussed in details.

Digital Archiving in the National Archives of Hungary

The currently used archiving system in the National Archives of Hungary was launched in 2013. It was developed according to the OAIS Model and it covers all functionalities of it. The system handles the archived content as information packages. For archiving relational databases SIARD the archive uses the SIARD format.

The Regulation on the Procedures and Technical Requirements for the Submission of Born Digital Public Files was introduced in 2016, which determines, that the transfer must carried out in the form of OAIS compliant information packages

Citations

[1] https://public.ccsds.org/pubs/650x0 m2.pdf

[2] Anderson et al., 2016, p. 80 (https://indico.cern.ch/event/666320/attachments/1641822/2645929/PV2018-RAL-CONF-2018-001.pdf)

[3] https://www.ech.ch/vechweb/page?p=dossier&documentNumber=eCH-0165&documentVersion=1.0

[4] https://www.ech.ch/vechweb/page?p=dossier&documentNumber=eCH-0165&documentVersion=1.0

[5] https://www.ech.ch/vechweb/page?p=dossier&documentNumber=eCH-0165&documentVersion=1.0

[6] www.eark-project.com

[7] http://www.eark-project.com/resources/conference-presentations/finconfpres/81-day-2-4-using-e-ark-to-preserve-relational-databases-in-hungary

[8] http://www.eark-project.com/resources/conference-presentations/finconfpres/81-day-2-4-using-e-ark-to-preserve-relational-databases-in-hungary

[9] Thirifays et al., 2016, (http://www.eark-project.com/resources/project-deliverables/91-d532)

[10] The new version of SIARD Suite, published in 09.07.2018 is already SIARD 2.0 compatible https://www.bar.admin.ch/bar/en/home/archiving/tools/siard-suite.html

[11] Currently supported RDBMS include Oracle, MySQL, PostgreSQL, Microsoft SQL Server and Microsoft Access

[12] https://www.database-preservation.com/

[13] Thirifays et al., 2016, (http://www.eark-project.com/resources/project-deliverables/91-d532)

References

Alex Thirifays, Zoltán Lux, Jože Škofljanec, Gregor Završnik, Anja Paulič, Anders Bo Nielsen, Phillip Tømmerholt, Janet Anderson, Richard Healey, Kuldar Aas, Andrew Wilson, Jan Aspenfjall, and David Anderson. D5.3 e-ark dissemination information package (dip) final specification. Technical report, E-ARK Project, 2016.

Janet Anderson, Miguel Ferreira, Richard Healey, Zoltán Lux, and Sven Schlarb. Database Archiving and Big Data Techniques from the E-ARK project. Science & Technology, Harwell, UK 15th-17th May, 2018 Esther Conway (editor), Kate Winfield (editorial assistant) PV2018: Proceedings of the 2018 conference on adding value and preserving data. 2018. p. 80-85. https://indico.cern.ch/event/666320/attachments/1641822/2645929/PV2018-RAL-CONF-2018-001.pdf

Je páska stále moderné médium pre archiváciu dát?

Stanislav Dzúrik, IBM Slovakia, Bratislava, SR

Abstrakt
Spoločnosť IBM je dlhoročným lídrom v oblasti informačných technológií. Rovnako patrí medzi spoločnosti, ktoré sa najviac podieľajú na vývoji nových technológií v tejto oblasti. Jednou z nich je aj samotný vynález a nepretržitý vývoj páskových technológií počas obdobia viac ako šesťdesiatich rokov. Páskové knižnice určené pre veľkokapacitné dátové archívy využívajúce lineárne páskové technológie vyvinuté spoločnosťou IBM sú v súčasnej dobe používané ako dátové úložiská pre CDA UKB s vlastnosťami dlhodobého dôveryhodného dátového úložiska. Cieľom príspevku je oboznámiť auditórium o novinkách na poli lineárnych páskových technológií, ich ďalšom plánovanom vývoji v najbližšom období, prečo páskové médiá aj dnes zostávajú najbezpečnejším a cenovo najvýhodnejším nosičom pre dátové archívy a aké sú najnovšie možnosti tvorby archívov s ich využitím.

Abstract
IBM is leader in the area of information technologies for a long time. IBM belongs between companies with the most amount of patents per year and with its own research and development. One of them is invention and long term development of tape technologies during more than sixty years. Tape libraries dedicated to large data archives using linear tape technologies developed by IBM are used like data stores for CDA UKB. Goal of the article is to give overview of the news in area of the linear tape technologies, its further planned development in the next future, why tape media also today remains the most secure and price most effective medium for data archives, what are the newest possibilities of design of archives using tapes.

Úvod

Dlhodobé dátové archívy obsahujú dáta, ku ktorým sa nepristupuje príliš často, ale ich obsah je hodnotný a užitočný pre rôzne využitie. Takéto archívy často disponujú veľkým objemom dát v rozpätí jednotiek až desiatok PB. K základným požiadavkám pre takéto archívy patria vlastnosti ako je spoľahlivé dlhodobé uloženie dát bez možnosti ich zmeny, odolnosť archívu voči elektronickým alebo mechanickým poškodeniam dátových médií, jednoduché možnosti vytvárania ďalších kópií, zálohovania, obnovy alebo migrácie uložených dát v prípade potreby. Rovnako je dôležitý jednoduchý prípadne automatizovaný manažment a v neposlednom rade optimálne finančné nároky pri tvorbe prevádzkovaní alebo rozširovaní takéhoto archívu.

HW korekcia bitových chýb na páskovom médiu

V CDA archíve UKB sú využívané dva typy lineárnych páskových technológií LTO Ultrium a Jaguar. Je to práve technológia Jaguar, vyvíjaná spoločnosťou IBM, ktorá je využívaná práve na dlhodobú archiváciu dát. Ako je samotnou technológiou zabezpečená integrita uložených dát ? Dáta sú počas ukladania na médium okamžite čítané, pričom sa zisťuje ich konzistencia. V prípade zistenia nekonzistencie sa zápis zopakuje okamžite. Takýmto spôsobom sa zabráni ku vzniku chýb už pri prvom zápise. Medzi unikátne vlastnosti technológie Jaguar patrí aj extrémna odolnosť voči bitovým chybám a fyzickému poškodeniu médií. Pravdepodobnosť vzniku bitovej chyby je výrazne znížená generovaním dvoj úrovňového kódu na opravu chýb (Error Correction Code – ECC). ECC sa generuje jednak v smere prevíjania ako aj na šírku média zároveň, v rámci tzv. „sub data setu“. Ďalšou technológiou, ktorá výrazne znižuje pravdepodobnosť vzniku bitovej chyby a zároveň zvyšuje odolnosť voči poškodeniu média je tzv. „deep interleaving“. Dáta sú na páskovom médiu rozložené špecifickým spôsobom, a to tak, že pásková mechanika dokáže tolerovať výpadok dát v zvislom páse na médiu širokom až 11 mm a prečítať dátové sety aj v prípade výpadku dvoch lineárnych stôp z celkových šestnástich na médiu. Ak by sme si štatisticky porovnali počet nedetekovaných bitových chýb, tak pre „enterpise“ SATA disk pripadá jedna nedetekovaná bitová chyba na cca 10 PB dát a pre Jaguár páskové médium pripadá jedna nedetekovaná bitová chyba na cca 100 Exabyte, čo je až 10-tisíckrát menej. Páskové médium je teda z tohoto pohľadu ďaleko spoľahlivejšie médium pre dlhodobú archiváciu dát ako vysoko kapacitné disky.

Prevedenie páskových kaziet

IBM páskové kazety majú vysokokvalitné prevednie, kontinuálne vylepšované na základe dlhoročných skúseností v praxi. Sú dizajnované tak aby boli vysoko odolné voči mechanickému poškodeniu média. Kazety majú robustný dizajn a sú zhotovené z pevného hustého plastu. Každé médium je testované na funkčnosť a aj na voľný pád z výšky až dvoch metrov (jedného metra v prípade LTO médií). Kazety umožňujú bezpečné a rýchle založenie páskového média do páskovej mechaniky ako aj jeho rýchle prevíjanie pri súčasnej eliminácii jeho poškodenia.IBM poskytuje na páskové médiá až päťročnú záruku a garantuje tridsaťročnú životnosť archívnych dát.

Efektívne a bezpečné ukladanie dát

Páskové mechaniky majú integrovanú kompresiu dát. Kompresia sa aktivuje a vykonáva na HW úrovni páskovej mechaniky a nemá negatívny vplyv na dátovú priepustnosť zariadenia. Naopak s rastúcim kompresným pomerom sa dosahuje vyššia priepustnosť dát. Pred samotnou kompresiou dát zariadenie otestuje efektivitu ich kompresie. Ak sa ukáže, že kompresia by bola neefektívna, nevykoná sa. Typický dosiahnuteľný kompresný pomer sa pohybuje v rozpätí 1:2 až 1:3. Metadáta sú ukladané na páskové médium a sú chránené dvoj úrovňovým ECC.

Páskové mechaniky umožňujú aj kryptovanie dát na HW úrovni. Manažment kryptovacích kľúčov je možné vykonávať na úrovni samotnej páskovej knižnice alebo aplikácie. Samotná enkrypcia má len nepatrný vplyv na dátovú priepustnosť zariadenia. Páskové médiá môžu disponovať aj WORM (Write Once Read Many) funkcionalitou, ktorá zabraňuje modifikácii dát uložených na médiu. Táto vlastnosť je nutná pre krátkodobú ako aj dlhodobú archiváciu dát.

Páskové mechaniky a páskové knižnice

Spoločnosť IBM ponúka svojim klientom ucelené portfólio páskových knižníc. Ponuka začína jednoduchými zariadeniami typu autoloader a končí vysoko škálovateľnými páskovými knižnicami TS4500 dosahujúcimi kapacity až 180 PB pre dáta a s priepustnosťou až 46 GB/s bez použitia kompresie. Hustota uložených dát v GB/m² pri použití IBM páskových knižníc je dnes veľmi vysoká. Pásková knižnica IBM TS4500 dnes

poskytuje možnosť uloženia viac ako 8 PB dát na ploche 4,23 m², alebo viac ako 38 PB na ploche 8,36 m² natívne, bez použitia kompresie. V porovnaní s rovnakou diskovou kapacitou je to polovičná veľkosť zabratej plochy v datacentre. Klienti, ktorí dnes využívajú páskové knižnice IBM TS3500 môžu využiť páskové mechaniky a dátové skrine v novej knižnici IBM TS4500 a tak ochrániť svoje investície z minulosti. Podobná možnosť ochrany investícií pri diskových poliach spravidla neexistuje.

Pásky majú vynikajúce TCO

Z pohľadu celkových nákladov na dlhodobú archiváciu dát je páska najekonomickejším riešením. Celkové náklady na jeden PB počas obdobia desiatich rokov dát môžu dosiahnuť v prípade použitia veľkokapacitných diskov až niekoľko miliónov EUR. V prípade použitia páskových médií ide o stotisíce EUR. Celkové náklady sú teda podstatne nižšie. Suma zahŕňa náklady spojené s prvotným nákupom a nákladmi na prevádzku (HW a SW podporou, nákladmi na miesto v datacentre, spotreba elektrickej energie na prevádzku a chladenie, manažment, ľudí a prác).

Vízia budúcnosti páskových médií

Aká je ďalšia perspektíva páskových technológií? Niektorí výrobcovia už pred niekoľkými rokmi tvrdili, že páskové zariadenia sú „mŕtve“ a stratili svoj význam. Život však ukázal, že nemali pravdu, čo je zrejmé už z uvedených faktov. V komerčnej sfére dnes dokážeme na jedno archívne páskové médium uložiť 15TB dát bez použitia kompresie. Prenosová rýchlosť jednej páskovej mechaniky je 360 MB/s. Nie je jednoduché dosiahnuť takýto dátový tok z relatívne lacných vysokokapacitných pevných diskov. Materiál dnes používaný na páskové média je BaFe (Barium Ferrit). Tento materiál je hladší, ako v minulosti používaný materiál zhotovený z tradičných kovových častí. BaFE umožňuje zápis s veľkou bitovou hustotou a je vysoko odolný voči demagnetizácii.

Vďaka tomu postačuje prevíjať páskové médiá raz za rok a garantujeme, že dáta vydržia na archívnom médiu až 30 rokov. Technické detaily o BaFe si môžete pozrieť na tomto web linku:

http://www.fujifilmusa.com/products/tape_data_storage/innovations/barium_ferrite/index.html

2. 8. 2017 IBM Research dosiahla nový rekord v ukladaní dát na pásku. IBM laboratóriá v Zurichu v spolupráci s našim partnerom pre vývoj médií, spoločnosťou Sony, demonštrovala nový svetový rekord – schopnosť uložiť dáta na páskové médium s hustotou až 1296Gbit/cm². Týmto bol prekonaný posledný rekord z nedávnej minulosti – 793Gbit/cm². Toto až 63% vylepšenie znamená, že celkovo bude možné uložiť až 330TB (nekomprimovaných) dát na jednu páskovú kazetu. IBM je presvedčená že dokáže zdvojnásobiť kapacitu páskovej kazety každým nasledujúcim rokom v priebehu ďalšej dekády. V ére "big data" a "cloud" to znamená, že páska zostane najlacnejším úložným médiom na planéte. Z pohľadu TCO sa nič nedotkne ekonomiky dátových pások. Páskové médiá zostávajú ideálnou platformou pre zálohovanie dát a rovnako sú ideálnym médiom pre dáta v dlhodobých archívoch. Bližšie informácie o tomto rekorde nájdete na web linku:

http://www.guidingtech.com/70544/little-capsule-can-store-330tb-data/?lipi=urn%3Ali%3Apage%3Ad_flagship3_feed%3Bn92qxTj6RrW8 moyhDPoDfw%3D%3D

V blízkej budúcnosti bude prekonaná natívna kapacita páskového média 18TB a prenosová rýchlosť dát jednej páskovej mechaniky 400 MB/s. To všetko pri použití BaFe média. Už v rokoch 2022 – 2023 sa očakáva, že sa na trh uvedú médiá s kapacitami 50 až 60TB. Momentálne sa v laboratóriách vyvíjajú média na báze Stroncium Ferit technológie. Kapacita týchto médií by mala dosahovať stovky TB.

Súčasné možnosti budovania dlhodobých archívov

Súčasným trendom v informačných technológiách sú softvérovo definované prípadne hyperkonvergované riešenia. Dôvodom je maximálne efektívne využitie HW zdrojov pre väčší počet aplikácií. Na týchto princípoch je dnes možné riešiť aj dlhodobé hybridné archívy. Umožňujú to technológie implementované v IBM Spectrum Scale a IBM Spectrum Archive softvéroch. Prvý z nich umožňuje vytvorenie hybridného viacúrovňového súborového systému s možnosťou nastavenia politík pre automatické ukladanie dát, politík súvisiacich presunom dát na iný typ média vzhľadom na frekvenciu ich použitia alebo vzniku nejakej udalosti. Zároveň umožňuje zamedziť modifikácii dát a umožňuje vymazanie dát v čase ich exspirovania. IBM Spectrum Archive softvér umožňuje pripojenie páskovej knižnice ako jedného z transparentných úložných vrstiev archívu. Teda dáta, ktoré je potrebné modifikovať alebo sa k nim často

pristupuje sú umiestnené na rýchlejšej úložnej vrstve a dlhodobo archivované dáta sú uložené na páskových médiách. Môžeme vytvárať kópie takéhoto archívu ako aj jeho zálohy. Kópia takéhoto archívu môže byť uložená na podobnom type hybridného archívu, alebo čisto na páskach alebo dokonca aj v cloud.

Zhrnutie

Páskové zariadenia a páskové médiá sú dnes cenovo najefektívnejšie a z pohľadu konzistencie a bezpečnosti uloženia archivovaných dát najvhodnejšie pre tvorbu dlhodobých archívov. V súčasnosti existuje silný komerčný plán vývoja (roadmap) páskových technológií či už na úrovni páskových knižníc, tak aj na úrovní páskových médií. Vďaka príchodu nových technológií pre archiváciu na báze SW definovaných riešení je možné plne automatizovať ukladanie a manažment dát v dlhodobých archívoch na základe nastavenia politík. Kópie moderných archívov je možné ukladať do cloud a tým eliminovať ďalšie riziká straty dát pri súčasnej optimalizácii nákladov.

The education of web archiving

László Drótos, Márton Németh, National Széchényi Library, Budapest, Hungary

Abstract
The article is focusing on three main issues. At first, an overview is being offered about an online research seminar for PhD students and web-archiving professionals organized by the NETLAB Research group, Aarhus University, Denmark. Secondly, the recently established Education and Training Working Group of the IIPC consortium is being introduced. A quick overview is being offered about a brief survey on best web archiving education practices and future. Thirdly, a Hungarian web-archiving training concept is being described. The training will be organized by the Library Institute for any kind of cultural heritage professionals that want to get basic skills and competences in this field.

keywords: education, web archiving, e-learning, NETLAB, IIPC

Introduction

Our paper is covering three main topics. Firstly, we are offering a short overview about an online seminar that was organized for PhD students and professional experts about web archiving by the NetLab research group, Aarhus University, Denmark. Secondly, the initial activities of the newly established Training Working Group of International Internet Preservation Consortium (IIPC) being introduced. A comprehensive survey made by this working group focused on best practices, current experiences and plans for the future. The first results are being presented. Thirdly we are introducing the initial plan and curriculum of a web archiving course in collaboration with the Library Institute. The course will focus on public collection professionals offering them an overview about tools and methodology of web archiving.

1. NetLab online course

The NetLab research group at Aarhus University, Denmark is a part of the national DIGHUMLAB[1] research infrastructure network led by prof. Niels Brügger. They have started to offer online courses about web archiving for two years. The major target groups are PhD students, public collection professionals and researchers. In the autumn of 2017 they introduced their first course entirely in English for an international audience[2]. The participation was free and it was really optimal to make new connections among the experts of the newly established web-archiving projects in Hungary and Belgium and work together with Danish colleagues as well. The online seminar was being held in a password protected Moodle-based e-learning interface. There we could access the exercises, training materials, hand-in the answers and make interactive conversations on the course forum. This interface was only available during the course but the participants could save all materials just after the finishing. The seminar covered five main topics A handbook by Janne Nielsen offered a general support through all topics and exercises.[3]

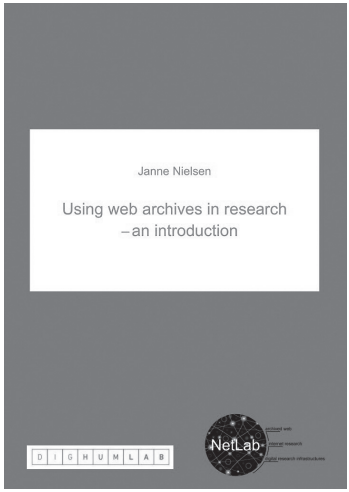


Figure 1 Cover of the course book

1 More details about the educational activities of NETLAB consortium, retrieved12.06.2018 http://www.netlab.dk/services/courses/

2 Course description, retrieved 12.06.2018. http://www.netlab.dk/wp-content/uploads/2017/04/NetLab-Web-Archiving-Course-Brochure.pdf

3 Freely available, retrieved 12.06.2018. http://www.netlab.dk/wp-content/uploads/2016/10/Nielsen_Using_Web_Archives_in_Research.pdf

At the beginning of the course the professional background, expectations to the course and the types of involvement with web-archiving were discussed among the participants. The pedagogic style of the seminar is constructivist. It heavily relies on the professional profiles and level of involvement of the participants. In each semester the profile of the seminar by certain groups is largely different on this way. Followed by an overview about general interests in web archiving of the members, the second task was to specify professional topics, tasks and formulate a small-scale research plan related to web archiving. The third task was to find three websites by your own interests (the sites must be in operation for minimum one year) and describe the archiving challenges of them. Samples could be found in Internet Archive[4] or any publicly available archived site from a national web archive could be also used. By the fourth task a collection strategy of websites had to be formulated. Followed by that the appropriate software background had to be selected and pilot harvests had to be initiated. Finally, the overall experiences had to be summarised in a report. This task appeared to be the most useful one because we could share experiences about pilot harvests with our Belgian colleagues and we could try to find answers about several archiving challenges. We could evaluate various software tools and analysing the harvest results. Our Danish colleagues from the Danish Web Archive[5] also could share with us their own experiences. The major goal was to formulate relevant questions about practical tasks in order to effectively formulate further of our own web-archiving pilot projects in Belgium and in Hungary. The final, fifth task was to make a general closing overview by discussions and filling up an evaluation survey made by the organizers. At the end all participants got an official certificate, demonstrating the completion of the course.

In case of this web-archiving seminar the applied pedagogical style turned to be really effective. The theoretical background was available in written form for individual studying. The lessons based on this theoretical core were really practice-based focusing on specific tasks and challenges. A major aim was to ensure the long-term application of course experiences on our own job in an effective way. We could learn the most from each other. By planning tasks, discussing software problems, archiving issues and resolving some challenges made us really valuable experiences. The course effectively helped the foundation of our own web-archiving pilot project in the National Széchényi Library.

4 Archive.org, retrieved 2018.06.12., http://www.archive.org

5 Netarkivet, retrieved 12.06.2018., http://www.netarkivet.dk

Figure 2 Course Certificate

2. IIPC Training Working Group (IIPC TWG)

Members of the IIPC international consortium are public and private organizations, institutions that are preserving online materials[6]. Primary tasks of the consortium are the development of technologies, methodologies, standards related to web archiving, sharing national best practices, supporting international collaboration, granting the broad access to the archived web materials and helping to re-use these datasets in various ways. The Training Working Group (TWG) had established at the end of 2017[7]. By their first project a survey was compiled[8]. The main aim was to collect basic information about national web archiving projects: Who, Where and in what kind of frameworks are working with web-archiving. The survey was also focused on human background of each institution and the aims and needs of professionals in education and training aspects of web archiving The survey was open in January, 2018. A quick summary of the results can be presented[9].

6 IIPC portal address, retrieved 12.06.2018 http://www.netpreserve.org

7 A description of Training Working Group is available on the following link, retrieved 12.06.2018
http://netpreserve.org/about-us/working-groups/training-working-group/

8 Survey link, retrieved 12.06.2018.,
https://www.surveymonkey.com/r/V7 MVXXW

9 The summary based on non-public IIPC working materials. Public references are not available.

The answer of respondents to the survey was 224, representing a global professional group from five continents. Web archiving activities have mainly done by universities, research institutes and in a smaller but relevant scale: national libraries. The number of archives with web archiving activities is also relevant furthermore we can find museums, audio-visual archives and some commercial actors in this field. The average number of people working with web archiving issues is really low. By the half of the institutions the respondents belonging to less than one full-time professional person is focusing on this issue. About a quarter of the respondents determined the number of people between 1 and 3. Nine percent of the institutions, organisations are working with at least 10 people on web archiving activities. The other institutions and organisations employ 3-5 people for these tasks. The third question focused on the type of activities related to web archiving. Most of the related people are curating content and setting up regulations, standards. Other main tasks are (by relatively the same weight): making metadata; quality assurance, communication tasks, harvest management. The least number of people in web-archiving field are the software developers. Most of the respondents have public collection background and only a small portion of them have relevant IT experience. Most people started to work with web-archiving tasks very recently. To put these tasks to the general service portfolio of a public collection appears to be a big challenge to them. Many of the respondents referred that they are planning to work with web-archiving in the future but they do not have any practical experience recently.

The next couple of question focused on the education and training aims in web archiving field. By the answers it appears that we are still at the very beginning of our professional way. Most of the responding people recently rely on online resources in order to develop their professional competences. The number of any kind of organised training activities is marginal. A relatively large number of respondents are currently without any kind of trainings. The least number of people are attending in courses by accredited curricula. Where any kind of training option is available it mainly focusing on workshops, formulated by informal frameworks or organized by some kind of professional organisation. Most of the respondents want to develop their web-archiving related competences in IT-field, by focusing on digital preservation standards, technologies and the education of use of relevant software tools. The most popular learning forms are webinars and some courses based on personal attendance.

The IIPC TWG has started to plan various training activities based on the survey experiences. The major aim is to effectively support the web archiving institutions and broad target groups of web-archiving related professionals. The second step was the collection of a list of trainings and courses by all of their major features that are already

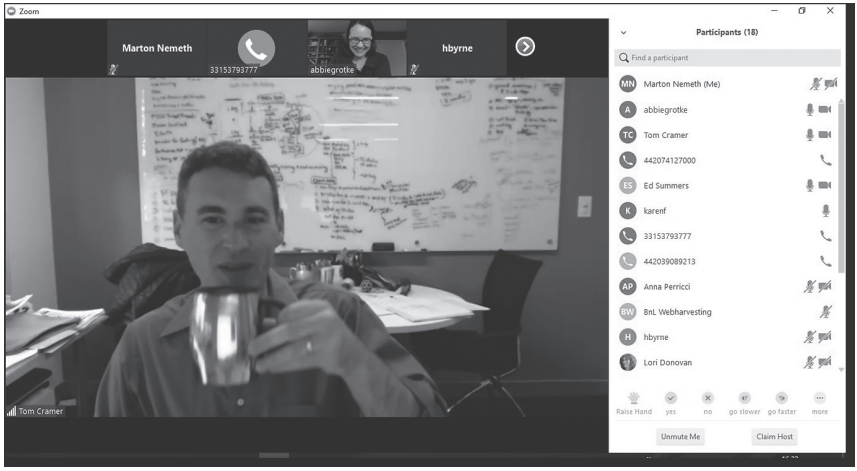
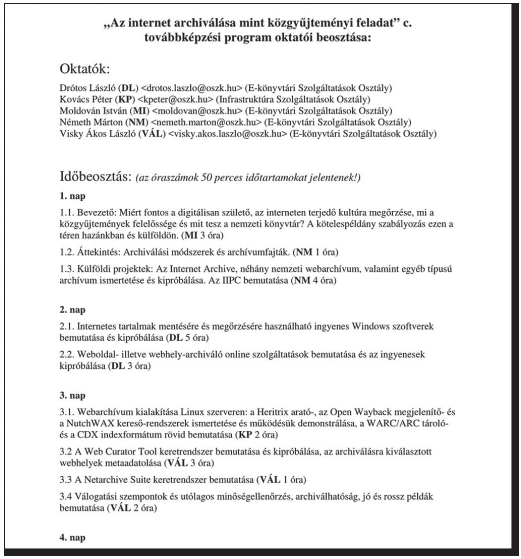


Figure 3. Collaborative work in an online meeting of IIPC Training Working Group

available in various countries, and learn form the best practices. The third step has been taken recently by starting to plan an online education environment that can be used by the IIPC members in general and can be adapted to each member's needs.

3. Training plans in Hungary

The National Széchényi Library has started a comprehensive project in order to establish a new national library system (OKR-project). As a segment of this large project, the web archiving pilot project has started at the beginning of 2017. By consulting public collection professionals, a definite aim has appeared to establish a 30 hour long special training to this target group. The key of success of a national web archiving model is an active collaboration within the public collection sector. The main goal of the special training therefore to introduce public collection professionals to the major technical background of preserving online content, offer an overview about international projects and present our own activities in this field. By completing this course, people should be able to create web archive collections in their workplace or for private purposes. They also have to be able successfully participating on the building process of a Hungarian Internet Archive by their own competences. Main target groups are librarians, archivists, museum professionals. The course will be offered by the web archiving team of the Electronic Library Department of the National Széchényi Library and by IT professionals from the Department of IT services.



„Az internet archiválása mint közgyűjteményi feladat" c. továbbképzési program oktatói beosztása:

Oktatók:

Drótos László (**DL**) <drotos.laszlo@oszk.hu> (E-könyvtári Szolgáltatások Osztály)
Kovács Péter (**KP**) <kpeter@oszk.hu> (Infrastruktúra Szolgáltatások Osztály)
Moldován István (**MI**) <moldovan@oszk.hu> (E-könyvtári Szolgáltatások Osztály)
Németh Márton (**NM**) <nemeth.marton@oszk.hu> (E-könyvtári Szolgáltatások Osztály)
Visky Ákos László (**VÁL**) <visky.akos.laszlo@oszk.hu> (E-könyvtári Szolgáltatások Osztály)

Időbeosztás: *(az óraszámok 50 perces időtartamokat jelentenek!)*

1. nap

1.1. Bevezető: Miért fontos a digitálisan született, az interneten terjedő kultúra megőrzése, mi a közgyűjtemények felelőssége és mit tesz a nemzeti könyvtár? A kötelespéldány szabályozás ezen a téren hazánkban és külföldön. (**MI** 3 óra)

1.2. Áttekintés: Archiválási módszerek és archívumfajták. (**NM** 1 óra)

1.3. Külföldi projektek: Az Internet Archive, néhány nemzeti webarchívum, valamint egyéb típusú archívum ismertetése és kipróbálása. Az IIPC bemutatása (**NM** 4 óra)

2. nap

2.1. Internetes tartalmak mentésére és megőrzésére használható ingyenes Windows szoftverek bemutatása és kipróbálása (**DL** 5 óra)

2.2. Weboldal- illetve webhely-archiváló online szolgáltatások bemutatása és az ingyenesek kipróbálása (**DL** 3 óra)

3. nap

3.1. Webarchívum kialakítása Linux szerveren: a Heritrix arató-, az Open Wayback megjelenítő- és a NutchWAX kereső-rendszerek ismertetése és működésük demonstrálása, a WARC/ARC tároló- és a CDX indexformátum rövid bemutatása (**KP** 2 óra)

3.2 A Web Curator Tool keretrendszer bemutatása és kipróbálása, az archiválásra kiválasztott webhelyek metaadatolása (**VÁL** 3 óra)

3.3 A Netarchive Suite keretrendszer bemutatása (**VÁL** 1 óra)

3.4 Válogatási szempontok és utólagos minőségellenőrzés, archiválhatóság, jó és rossz példák bemutatása (**VÁL** 2 óra)

4. nap

Figure 4. An excerpt from the preliminary course-plan (in Hungarian)

Course curriculum consists of the following main modules:

- Getting to know internet preservation terminology, definitions and models in a basic theoretic framework.
- Get competences in a basic level of using some Windows-based archiving software, online services, and other useful software tools to build-up and support an archiving workflow.
- Get basic competences in a user level to the workflow and major components of a Linux-based web archive.
- Basic competences on the curation of web materials and major tasks for metadata enrichment of the archived material.
- Introducing web archives as a research subject. A basic overview about using web archives for research purposes. Foundations of planning and managing user-centred (mainly scientific) services based on web archived materials. Major competences related to create and maintain appropriate conditions of long-term sustainability of web-archives...

Followed by the accreditation period and granting appropriate funds the course can be started at the late autumn of 2018 in our hope. Besides this course we have started to plan an online course based on a blended learning-based curriculum. Course would be

completed partly online and partly by personal attendance. We hope that it also will be available at the end of 2018 by the latest for all people that are interested in long-term preservation of internet content.

Epilogue

In our paper an overview has offered about the structure and outcomes of an online professional course about web archiving that has managed from Denmark. We also offered a summary about the preliminary plans and basic activities of the IIPC Training and Working Group that offered us a major overview about the current framework, background and status of web archiving activities throughout the world. Last but not least we elaborated our training plans in Hungary. It is vital to train people with certain competences in order to build-up a national web-archive network. Based on this collaborative framework archiving activities can be done ordinarily and efficiently. A major pre-condition of the establishment of a well-functioning national network is to guarantee permanent professional development (both individually and on institutional level). Accredited trainings must be offered for web-archiving professionals in a permanent way to constantly keep their knowledge on a required level.

Bibliography

IIPC Training Survey Call, retrieved: 12.06.2018
https://netpreserveblog.wordpress.com/2017/12/14/iipc-training-survey/

IIPC Training Survey, retrieved: 12.06.2018
https://www.surveymonkey.com/r/V7 MVXXW

IIPC Training Working Group portal, retrieved: 12.06.2018
http://netpreserve.org/about-us/working-groups/training-working-group/

NIELSEN Janne*, Using the Web archives in Research (Theoretical course book of NetLab web archiving course),* retrieved: 12.06.2018
http://netlab.dk/wp-content/uploads/2016/10/Nielsen_Using_Web_Archives_in_Research.pdf

Website of the Danish Web Archive, retrieved: 12.06.2018
http://www.netarkivet.dk

A brochure of NetLab web archiving course retrieved: 12.06.2018
http://netlab.dk/wp-content/uploads/2017/04/NetLab-Web-Archiving-Course-Brochure.pdf

Website of the NETLAB web archiving course, retrieved: 12.06.2018
http://netlab.dk/services/courses/

Dlhodobé uchovávanie slovenského archívu digitálnych prameňov

Andrej Bizík, Univerzitná knižnica v Bratislave, Bratislava, SR

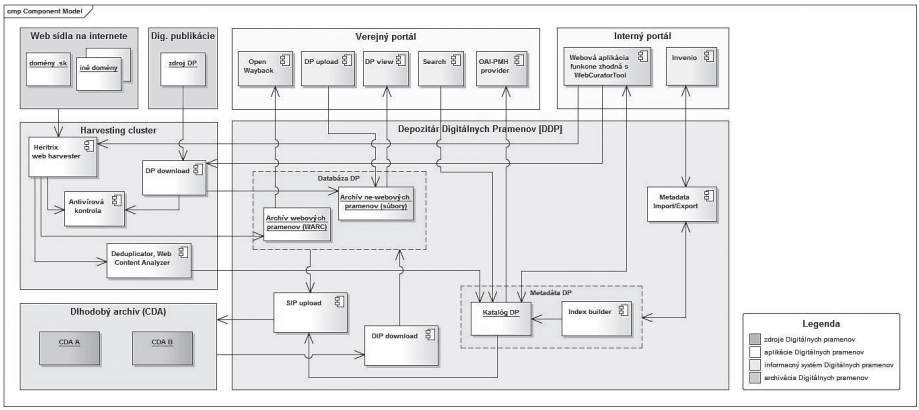
Abstrakt
Elektronické dokumenty a webový archív je treba „zakonzervovať“ podobne ako fyzické objekty trvalej hodnoty; realizuje sa to pomocou platformy dlhodobého úložiska. Počas celého procesu prenosu a uloženia dokumentov sa vyskytujú rôzne problémy a nástrahy. Samotné dlhodobé uloženie je komplikovaná operácia, ktorej ukážkou je sofistikované riešenie Depozitu digitálnych prameňov. Príspevok sa venuje dlhodobému uchovávaniu slovenského archívu digitálnych prameňov a popisuje problémy, ktoré sa pri tejto činnosti vyskytli. Integrovaná tvorba SIP (Informačný balík pre vklad – Submission Information Package) balíkov v Depozite digitálnych prameňov sa ukazuje ako rýchle a efektívne poloautomatické riešenie od vytvorenia až po uloženie SIP do Centrálneho dátového archívu (CDA). Skript prakticky vytvorí balíky, podpíše ich, uloží na dočasné úložisko, kde čakajú na potvrdenie kurátorom. Balíky označené trvalým identifikátorom sa v určitých časových úsekoch prenášajú na dočasné úložisko CDA, kde čakajú na spracovanie. Jednoduchá obsluha vyžaduje minimálny zásah kurátora, ktorý má odosielaný obsah neustále pod svojou správou.

Abstract
Electronic documents and web archives should be „preserved“ similarly to physical objects of lasting value; it is realized using the long-term storage platform. Throughout the process of transfering and storing documents, there are various problems and pitfalls. Long-term storage itself is a complicated operation, exemplified by a sophisticated solution Deposit of Digital Resources. The contribution focuses on the long-term storage of the Slovak archive of digital sources and describes the problems encountered in this activity. The integrated creation of SIP (Submission Information Package) packages in the Deposit of Digital Resources is a fast and efficient semi-automatic solution from creation to SIP storage in the Central Data Archive (CDA). The script will practically create packages, sign them, save them in a temporary repository, where they wait for confirmation by the curator. Permanent identifier packages are trans-

ferred to a temporary CDA repository for specific periods of time, waiting for processing. Simple operation requires minimal intervention by the curator who has the content sent constantly under his management.

Úvod

Cieľom projektu Digitálne pramene – webharvesting a archivácia e-Born obsahu bolo vybudovanie odpovedajúcej technickej, aplikačnej a organizačnej infraštruktúry na systematický zber a dlhodobú archiváciu slovacikálnych webových publikácií a pôvodného elektronického obsahu. Oficiálny rámec zadania určil len základné formálne ukazovatele projektu: počet informačných a dokumentačných databáz, počet nových technických zariadení, počet nových elektronických služieb on-line a napokon počet novovytvorených pracovných miest. Samotné zadanie pre Informačný systém Digitálne pramene (IS DIP) predstavovalo komplexný súbor funkčných a nefunkčných požiadaviek na architektúru, prieskum WWW, záber, kvalitu a kompletnosť, reporting, katalóg, sprístupňovanie, dlhodobú archiváciu a správu systému. Požadovaná funkcionalita sa odrazila v architektúre systému, ktorá pozostáva z autonómnych, navzájom prepojených funkčných blokov: verejného portálu, interného portálu, zberového clustera a repozitára. Dlhodobá archivácia sa rieši napojením na systém CDA [1].

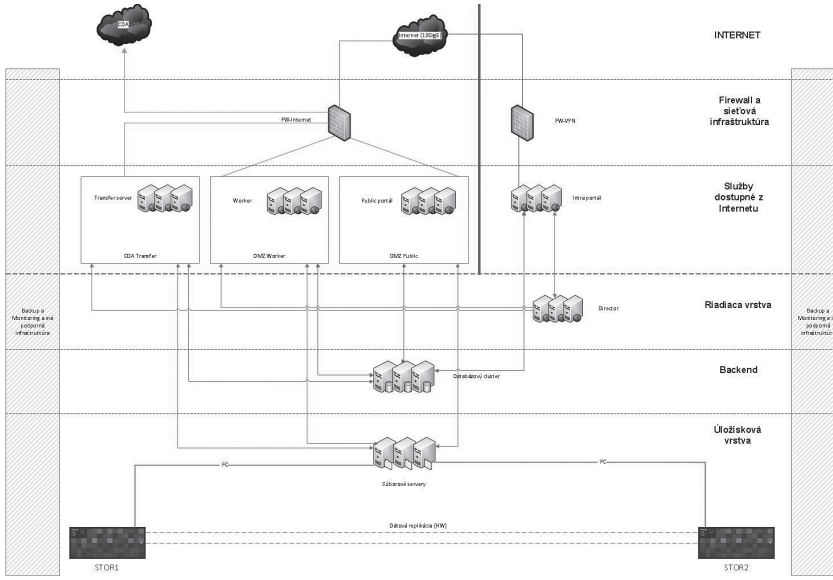


Obrázok 1: Funkčná schéma IS DDP

Technické riešenie

Technické riešenie IS DIP (Obrázok 2) je optimalizované na paralelný zber, spracovanie a indexáciu webového obsahu. Obsahuje všetok softvér potrebný na zber a transformáciu webovej stránky na archívny balík vo formáte WARC.

Systém disponuje tromi fyzickými servermi v clusteri, ktoré slúžia na riadenie aplikačných databáz, zdieľanie údajov a ukladanie archívnych balíkov a metadáta k zozbieraným dátam. IS DIP je vybavený 800 TB diskovým subsystémom pre potreby spracovania, indexácie a archivácie, ktorý výhľadovo spĺňa požiadavky na pracovný a úložný priestor vrátane potrebnej redundancie. Na konci tejto cesty sa nachádza prepojenie s CDA [1].



Obrázok 2: Technická schéma IS DDP

Proces generovania SIP balíkov

SIP balíky sa generujú samostatne pre Elektronické dokumenty a webový archív. Do SIP balíkov sa postupne zaraďujú úspešne vložené manifestácie a zbery domén zapísaním jedinečného „id", ktorý sa pomocou skriptov automaticky generuje podľa nastavených konfiguračných premenných.

Na strane DDP sa generujú identifikátory SIP balíkov, ktoré plne vyhovujú požiadavkám na ich generovanie popísané v dokumente Generická dohoda – Príručka Vkladateľa CDA[1]. SIP balíky obsahujúce manifestácie majú prefix EB (e-Born) , SIP balíky obsahujúce zbery domén majú prefix HA (harvest). Číselná časť identifikátora SIP balíka je odvodená z id záznamu v databázovej tabuľke SIP balíka.

Zbery domén, ktoré sú závislé na predošlých zberoch (z dôvodu deduplikácie, veľkosti a pod.), sú zaraďované do balíku SIP až po spracovaní SIP balíkov, ktoré obsahujú predošlé zbery domény. Do súboru mets-md.xml sa uvedie identifikátor AIP balíka, uloženého v CDA, obsahujúceho spracovaný predošlý zber domény.

Formát SIP balíka

Systém DDP generuje balíky vo formáte ZIP. Obsah balíka je v súlade s požiadavkami uvedenými v Príručke vkladateľa CDA[1]. Pre každý zber alebo manifestáciu vygenerované SIP balíky v hlavnom adresári obsahujú jeden alebo viac podadresárov. Názvy podadresárov generuje systém automaticky. Podadresáre, okrem súborov obsahujúcich samotný obsah manifestácii a zberov, obsahujú aj popisné metadáta vo formáte MARC v súbore marc.xml. Štruktúra názvov podadresárov je nasledovná:

- zbery (napr. www_stranasms_sk_.499996384):
  *www_stranasms_sk_* – URL zberu s nahradením bodiek a lomítok podčiarkovníkmi,
  499996384 – identifikátor zberu (id z databázovej tabuľky harvestov).
- e-Born – serialy (napr. Rok_2003_c_8_12.474479359):
  Rok_2003 – rok vydania,
  c_8_12 – číslo,
  474479359 – identifikátor manifestácie (id z databázovej tabuľky manifestácií).

1 Dostupná na adrese : http://cda.kultury.sk/?q=node/30

- e-Born – monografie (napr. CYBERCRIME.494360284):
  CYBERCRIME – titul monografie,
  494360284 – identifikátor manifestácie (id z databázovej tabuľky manifestácií).

Proces zálohovania

CDA prijíma od PFI informačné SIP balíky vyhovujúce Dohode o zverení obsahu na dlhodobú archiváciu v systéme CDA. SIP má pridelený identifikátor balíka vygenerovaný systémom Digitálnych prameňov. SIP balíky sú generované automaticky a sú ukladané do dočasného adresára na serveroch Digitálnych prameňov, kde čakajú na potvrdenie odoslania. SIP balíky sú digitálne podpísané certifikátom dodaným certifikačnou autoritou CDA.

Použitím webovej služby[2] v systéme CDA systém založí objednávku na vklad pre každý vygenerovaný balík. Po úspešnom založení objednávky je SIP presunutý prostredníctvom Wdav-protokolu do adresára pre odoslanie do CDA. Wdav je dostupný na url https://pfi.cda-a.kultury.sk/ddp-wdav a https://pfi.cda-b.kultury.sk/ddp-wdav. Prihlasovanie do Wdav-u vyžaduje dvojfaktorovú autentifikáciu certifikátom a používateľským menom a heslom. Odosielanie SIP do CDA je automatické, prostredníctvom internetovej siete.

SIP na strane CDA je skontrolovaný, či zodpovedá objednávke na vklad. Následne je spracovaný a trvalo uložený v CDA. Proces spracovania balíka môže trvať značný čas z dôvodu veľkosti balíka ale najmä z dôvodu dostupnosti kapacít spracovania na strane CDA.

IS DIP sa dotazuje v časových intervaloch na stav spracovania na strane CDA prostredníctvom webovej služby. V prípade úspešného spracovania uloží doručený identifikátor AIP do databázy a označí záznam o SIP ako ARCHIVED.

Konfiguračné premenné

Konfigurácia je získavaná aplikáciami z konfiguračného servera. Konfiguračný server poskytuje konfiguračné premenné podľa zadanej aplikácie, profilu a popisu. Konfigu-

2 Dostupná na adrese : https://pfi.cda.kultury.sk/pfi-webapp/ SubOrderWebService

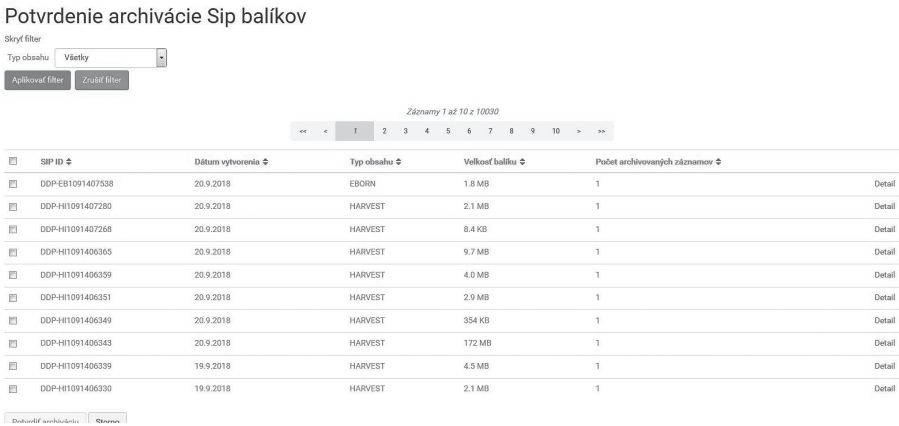
račné premenné pre generovanie SIP balíkov sa nachádzajú v súbore application.yml, ktorý je prepisovaný profilovým súborom. Application.yml obsahuje tieto hlavné premenné:

- cdaSipPrefix – Skratka PFI dohodnutá v Dohode.
- cdaProfile – Aktuálny profil pre vklad.
- subOrderWebService.url – URL na službu CDA pre vytvorenie objednávky na vklad.
- subOrderWebService.keyStore – Cesta na úložisko kľúčov vo formáte pkcs12[3], ktoré obsahuje kľúčový pár pre SSL autentifikáciu.
- subOrderWebService.passphrase – Heslo k úložisku kľúčov pre SSL autentifikáciu.
- subOrderWebService. verifyServerIdentity – Príznak povolenia verifikácie certifikátu servera.
- subOrderWebService. userName – Prihlasovacie meno systémového používateľa, ktorý má právo na volanie webovej služby CDA pre vklad objednávok.
- subOrderWebService. password – Heslo systémového používateľa pre volanie webovej služby CDA pre vklad objednávok.
- tmpDir – Absolútna cesta k adresáru pre ukladanie vygenerovaných SIP balíkov pred vytvorením objednávky v CDA.
- harvest.archiveMaxSize – Maximálna veľkosť SIP balíka v GB.
- harvest.archiveMaxObjects – Maximálny počet objektov v SIP balíku.
- harvest.archiveMinSize – Minimálna veľkosť SIP balíku v GB.
- eborn.archiveMaxSize – Maximálna veľkosť SIP balíku v GB.
- eborn.archiveMaxObjects – Maximálny počet objektov v SIP balíku.
- eborn.archiveMinSize – Minimálna veľkosť SIP balíku v GB.
- signature.file – Cesta na úložisko kľúčov vo formáte pkcs12[3], ktoré obsahuje kľúčový pár pre podpisovanie SIP balíkov.
- signature.sipSigPassword – Heslo k úložisku kľúčov pre podpisovanie SIP balíkov.

Potvrdenie archivácie Sip balíkov

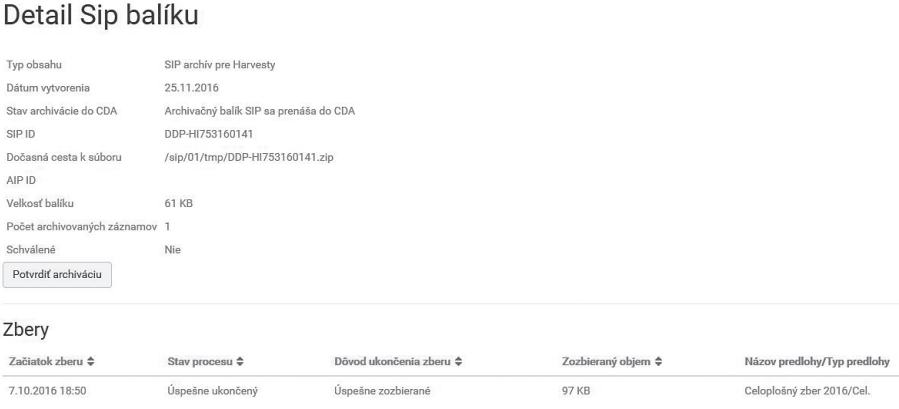
Zobrazený formulár obsahuje zoznam SIP balíkov, ktoré ešte neboli schválené na odoslanie do CDA. Filter umožňuje zobrazenie všetkých balíkov pre e-Borny alebo Harvesty.

3 *Špeciálny súbor pre digitálny certifikát.*



Obrázok 3: Potvrdenie archivácie SIP

Kurátor označí balíky pre export do CDA pomocou výberového tlačidla na začiatku riadku. Formulár umožňuje označiť viacero SIP balíkov a hromadné označenie pomocou výberového tlačidla v riadku hlavičiek stĺpcov. Po použití tlačidla „Potvrdiť archiváciu“ budú predmetné balíky presunuté na súborovom systéme do priečinka, z ktorého budú automaticky odoslané do CDA a vytvorí sa objednávka na vklad do CDA (viď používateľská príručka CDA pre PFI). Potvrdenie archivácie je nevratný proces.



Obrázok 4: Detail SIP balíka

Obrázok 5: Detail obrazovky archívny záznam

Pred odoslaním balíka do CDA je vhodné vykonať zbežnú kontrolu SIP použitím tlačidla detail na konci riadku predmetného záznamu. Z detailu SIP balíka je možné zobraziť detail archívneho záznamu.

Informácia o úspešnom spracovaní SIP na strane CDA je okrem PFI modulu na stránkach CDA dostupná aj v detaile Zberu domény alebo pre elektronický dokument na obrazovke Archívny záznam.

Doteraz zaznamenané chyby a problémy

Po odoslaní balíkov SIP do CDA nasleduje validácia dát. Chybné dáta/balíky, ktoré neprejdú sa analyzujú. Časté chyby ktoré sa vyskytli:

- Diakritika, nesprávny formát názvu súboru, názov súboru obsahoval dĺžne a iné. Príklad nevyhovujúci názov súboru:
  Nevyhovuje obmedzeniam daných dohodou (regulárny výraz: [A-Za -z0-9()+,-.:=@;$_!‘%/?#]+) ./www_parlamentnelisty_sk_tagy_mesezni-kov.459521785/marc.xml*
- Špeciálny znak & – znak „&“ nahrádza niektoré znaky (ý za &#37), no v niektorých prípadoch chybne nahradil znak, ktorý vyhodnotila kontrola ako chybu diakritiky.

- Bol detegovaný vírus – archív môže obsahovať aj nechcený alebo nebezpečný obsah.
- METS nie je validný xml dokument – súbor mets.xml je chybný, treba analyzovať obsah a jednotlivo poopravovať zistené nedostatky.

Záver

Príspevok zobrazuje vkladanie SIP balíkov do CDA ako automatizovaný proces, ktorý nemusí byť vôbec zložitý. Na pozadí je množstvo samostatných procesov od vytvorenia, odoslania, prijatia, výberu až po potvrdenie o uložení. Celý proces vkladu je vo výsledku pre kurátora jednoduchá aktivita, ktorá má za sebou množstvo plánovania a činností. V budúcnosti nasledujú výbery veľkých dát, zaobalených v AIP balíkoch, ktoré bude treba spoľahlivo prevziať v celku so všetkými naviazanými balíkmi. Táto aktivita je v začiatkoch, keďže zatiaľ nebolo potrebné staršie dáta mazať pre dostatočnú kapacitu. Odosielanie SIP do CDA plánujeme priamym prepojením serverov, pomocou optickej siete. Priamym prepojením zrýchlime odosielanie a prijímanie všetkých archivovaných údajov. Dlhodobé uloženie archívu Depozitu digitálnych prameňov takto spĺňa cieľ dlhodobej ochrany elektronických dokumentov a webových stránok, ktoré sú vzhľadom na svoju nehmotnú povahu krehkou a ohrozenou súčasťou nehmotného kultúrneho dedičstva.

Použité skratky

WARC – Web archive format – webový archívny formát.
CDA – Digitálny archív s celoslovenskou pôsobnosťou, prevádzkovaný UKB.
DB – Dátová báza – množina štruktúrovaných dát.
Harvest – Zber digitálnych údajov.
Webharvest – Zber digitálneho obsahu z Internetu.
Metadáta – Štruktúrované informácie o primárnych dátach.
PFI – Pamäťová a fondová inštitúcia.
SIP – Submission Information Package (Informačný balík pre vklad).
AIP – Archival Information Package (Archívny informačný balík).
mets-md.xml – popisný súbor balíka pre archiváciu.

Dohoda – Dohoda o zverení obsahu na dlhodobú archiváciu v systéme CDA.
WebDAV – protokol pre vytváranie, zmenu a presun súborov medzi serverom a klientským počítačom.
e-Born – dokumenty, publikácie a informačné pramene pôvodne vytvorené v elektronickej forme (vzniknuté elektronicky).

Zoznam bibliografických odkazov

[1] ANDROVIČ, Alojz, Andrej BIZÍK, Peter HAUSLEITNER, Beáta KATRINCOVÁ, Iveta LACKOVÁ a Jana MATÚŠKOVÁ. Digitálne pramene – národný projekt zberu a archivácie v roku 1. *Knihovna Plus* [online]. Národní knihovna ČR. 2017, č. 1. ISSN 1801-5948 [cit. 25. septembra 2018]. Dostupné na internete: http://knihovnarevue.nkp.cz/kplus-web/archiv/2017-01/historie-a-soucasnost/digitalne-pramene-2013-narodny-projekt-zberu-a-archivacie-v-roku-1.

Diseminácia uložených archivovaných údajov z pohľadu zachovania dlhodobej ochrany archívu

Jaroslav Zeman, Univerzitná knižnica v Bratislave, Bratislava, SR

Abstrakt
Centrálny dátový archív (CDA) musí poskytovať aj možnosť výberu archivovaných dát. Tento príspevok objasňuje dôvody a metodický prístup k riešeniu uvedenej problematiky hromadnej diseminácie v prostredí CDA. Na ukážke realizovaného výberu pre potreby Národného osvetového centra vyžiadaných archívnych balíčkov múzea SNP vysvetľuje všetky úskalia spojené s diseminációu z pohľadu ochrany páskových nosičov pred nadmerným opotrebovaním. Cieľom uvádzanej metodiky je optimalizácia spracovania diseminačnej kampane z hľadiska výkonnosti a efektívnosti využitia páskových mechaník.

Abstract
The Central Data Archive (CDA) must also provide the possibility of selecting archived data. This paper explains the reasons and methodological approach to addressing the issue of mass dissemination in the CDA environment. Using the requested archive packages of Museum of the Slovak National Uprising made for the needs of the National Edification Centre as an example, the selection explain all the pitfalls associated with dissemination from the point of view of protection of strap carriers from excessive wear. The aim of the presented methodology is to optimize the processing of the dissemination campaign in terms of efficiency and effectiveness of the use of tape mechanics.

Úvod

Po trojročnej činnosti CDA sa okrem vkladov digitalizovaných diel a pamiatok začali čoraz častejšie vyskytovať požiadavky od pamäťových a fondových inštitúcií (ďalej

PFI) aj na ich výber z archívu. Pri tejto činnosti však dochádzalo k nekontrolovanému prístupu na pamäťové médium – magnetickú pásku. Napriek tomu, že používame pre uskladnenie najkvalitnejšie dostupné magnetické médiá typu Jaguár, tak aj tie majú stanovené podmienky pre dlhodobé uchovanie zapísaných informácií. Jedným z ukazovateľov je aj garantovaný počet vkladaní do páskovej mechaniky a tiež počet pretočení pásky v mechanike.

Práve toto je jeden z problémov, na ktorý poukazoval náš dodávateľ archívneho riešenia, ktorý nám zároveň aj zabezpečuje podporu pri prevádzkovaní systému. Počas výberu väčšieho počtu archívnych balíkov dochádzalo k náhodnému prístupu k týmto súborom na páskovom médiu, čo malo za následok takmer 100% počet vkladaní a pretáčaní magnetickej pásky pri každom súbore. Takto by sme výrazne skrátili životnosť magnetických pások a jedna z kópií uloženého obsahu by bola ohrozená.

Preto sme v spolupráci s dodávateľom, ktorý nám sprístupnil príkazy archívneho systému IBM, pracovali na metodike optimalizovaného výberu uložených súborov podľa uloženia na jednotlivých páskových médiách.

Vytvorenie postupu optimalizovanej diseminácie

Základný princíp úspešnej optimalizovanej diseminácie je dostať údaje z magnetickej pásky v takom poradí, ako sú za sebou zapísané na magnetickej páske. Takto treba postupne prečítať všetky pásky, na ktorých sa nachádzajú žiadané súbory.

Náš archivačný systém je postavený na hierarchickom úložnom systéme, ktorý na prvej úrovni tvorí diskové pole o kapacite 33TiB a na druhej úrovni potom pásková knižnica typu Jaguár. Pri vklade sa balíky najprv uložia do tohto diskového poľa a následne sa migračným procesom ukladajú na páskové médiá. Takéto riešenie je výhodné pre odstránenie závislosti od počtu páskových mechaník pri vklade tisícov balíčkov. Systém si sám riadi využitie mechaník, tak aby ich optimálne využil a bol výkonný.

Túto filozofiu sme museli zachovať aj pri procese výberu. Teda nechať systém rozhodnúť o spôsobe a poradí v akom ich sprístupní do prvej úrovne – diskového poľa. Kapacita diskového poľa je momentálne limitujúci faktor, ktorý určuje množstvo sprístupniteľných dát v jednom cykle. Preto musíme najprv vykonať analýzu požiadavky

na výber. Následne rozdeliť požiadavku na cykly do kapacity cca 30TiB. Potom realizujeme cyklus v štyroch fázach.

- Prvá fáza je zabezpečená zakázaním migrácie dát z diskového poľa na pásky.
- Druhá fáza vykonáva načítanie balíkov z pások do diskového poľa.
- Tretia fáza zahŕňa vytvorenie objednávok na diseminácia v aplikačnom rozhraní pre takto sprístupnené balíky. Po spracovaní sa uložia do webdavu alebo na magnetickú pásku typu LTO so súborovým systémom LTFS. Tým je zabezpečená kompatibilita pri výmene dát medzi inštitúciami PFI a CDA.
- Štvrtá fáza reprezentuje opätovné povolenie migrácie dát na pásky, kde migračné procesy overia prítomnosť vyžiadaných balíkov na páskach a z diskového poľa ich vymažú. Tým sa uvoľní priestor na diskovom poli a môže sa začať ďalší cyklus. Práve tieto obmedzenia kapacity diskového poľa sú príčinou nutnosti analýzy veľkosti požadovaných balíkov.

Riešenie problému získania informácií o balíkoch

Komplexná informácia o uložených archivovaných balíkoch je zapisovaná v katalógu. Ten sa intenzívne používa pri samotných operáciách vkladu, výberu a kontrole balíčkov. Preto sme museli hľadať iný zdroj informácií potrebných pre zadanie požiadavky na archívny systém IBM.

Analýzou dostupných databáz sme sa rozhodli využiť tabuľku jobdata, ktorá obsahuje informácie o prebiehajúcich úlohách a tiež poskytuje požadované informácie. Preto sme vyextrahovali relevantné polia z tejto tabuľky do jednoduchého textového súboru na serveri. Tento súbor sa pravidelne aktualizuje pridávaním riadkov o spracovaných vkladoch za ostatné obdobie. Takto vlastná príprava diseminácie nezaťažuje procesné databázy.

Druhým zdrojom informácie o použitých páskových médiách je získavaná podobným spôsobom z databázy DB2 použitím dotazovacích príkazov zálohovacieho systému IBM Spectrum Protect. Tu sú však informácie zapisované do samostatných textových súborov vytvorených pre každú pásku z dôvodu jednoduchej a rýchlej aktualizácie. Systém aktualizácie údajov o páskach je založený na skutočnosti, že na páskové médiá sa súbory = balíčky iba pridávajú a už sa nemažú. Preto stačí sledovať zoznam pások a ich status. Pásky v stave „FULL“ sú už uzavreté pre ďalší zápis a preto už nie je potrebné aktualizovať zoznam súborov na danom médiu. Aktualizujú sa iba informá-

cie na rozpracovaných médiách a na čerstvo uzavretých páskach, ktoré boli pri minulej aktualizácii ešte v stave rozpracovanosti „FILLING“.

Riešenie analýzy požiadavky na výber balíkov

Vytvorenie základného plánu pre diseminácia je riešené vlastným skriptom v jazyku bash, ktorému treba zadať ako parameter meno textového súboru so zoznamom archívnych balíkov (AIP). Ten je potrebné najprv vytvoriť z dodaného zoznamu od PFI. Výstupom skriptu je tiež textový súbor obsahujúci už požadované informácie pre ďalšie spracovanie. Obsah je zároveň zotriedený podľa názvu pásky a teda predpripravený na rozdelenie do cyklov, ak je to pre objem dát potrebné.

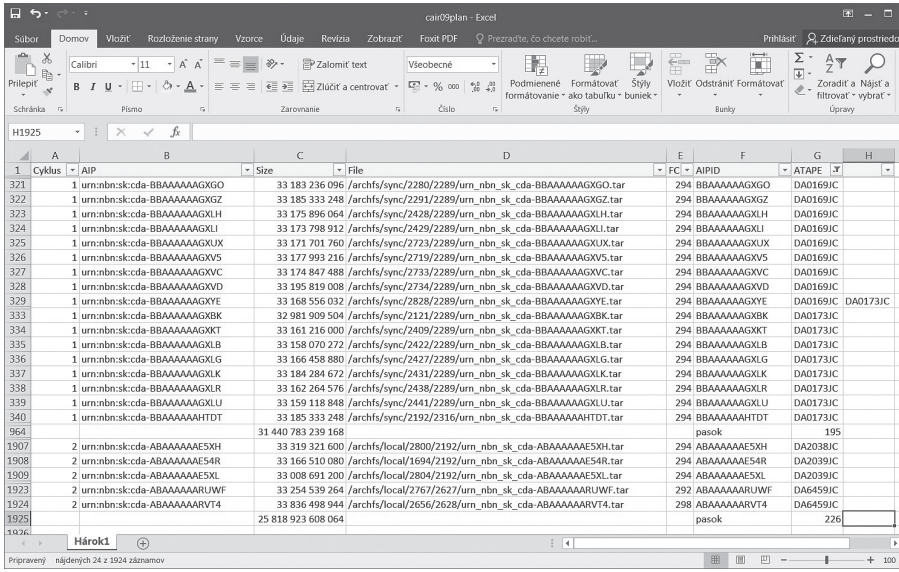
Súbor sa odošle ako príloha mailu aplikačnému administrátorovi. Ten si ju uloží lokálne na pracovnej stanici a následne importuje do Excelu alebo podobného tabuľkového programu ako textový súbor s poľami oddelenými medzerou.

Administrátor sa rozhodne o spôsobe spracovania na základe analýzy tejto tabuľky. Smerodajné sú informácie o sumárnej veľkosti požadovaných súborov a rozmiestnení týchto súborov na páskach, tak aby sa jedna páska nemusela načítavať vo viacerých cykloch ale iba v jednom. Toto je nevyhnutné pri požiadavkách presahujúcich kapacitu 30TiB.

Popis jedného cyklu optimalizovanej diseminácie

1. Fáza blokovania migračného procesu

Je zabezpečená vytvorením podadresára disem_lock na dohodnutom mieste diskového filesystému. Ten slúži ako semafor. Po spustení migračného procesu skript vyhodnotí existenciu podadresára s príponou „_lock“ a na základe tejto skutočnosti ihneď skončí vykonávanie ďalších príkazov.



	A	B	C	D	E	F	G	H
1	Cyklus	AIP	Size	File	FC	AIPID	ATAPE	
321	1	urn:nbn:sk:cda-BBAAAAAAGXGO	33 183 236 096	/archfs/sync/2280/2289/urn_nbn_sk_cda-BBAAAAAAGXGO.tar	294	BBAAAAAAGXGO	DA0169JC	
322	1	urn:nbn:sk:cda-BBAAAAAAGXGZ	33 185 333 248	/archfs/sync/2291/2289/urn_nbn_sk_cda-BBAAAAAAGXGZ.tar	294	BBAAAAAAGXGZ	DA0169JC	
323	1	urn:nbn:sk:cda-BBAAAAAAGXLH	33 175 896 064	/archfs/sync/2428/2289/urn_nbn_sk_cda-BBAAAAAAGXLH.tar	294	BBAAAAAAGXLH	DA0169JC	
324	1	urn:nbn:sk:cda-BBAAAAAAGXLI	33 173 798 912	/archfs/sync/2429/2289/urn_nbn_sk_cda-BBAAAAAAGXLI.tar	294	BBAAAAAAGXLI	DA0169JC	
325	1	urn:nbn:sk:cda-BBAAAAAAGXUX	33 171 701 760	/archfs/sync/2723/2289/urn_nbn_sk_cda-BBAAAAAAGXUX.tar	294	BBAAAAAAGXUX	DA0169JC	
326	1	urn:nbn:sk:cda-BBAAAAAAGXV5	33 177 993 216	/archfs/sync/2719/2289/urn_nbn_sk_cda-BBAAAAAAGXV5.tar	294	BBAAAAAAGXV5	DA0169JC	
327	1	urn:nbn:sk:cda-BBAAAAAAGXVC	33 174 847 488	/archfs/sync/2733/2289/urn_nbn_sk_cda-BBAAAAAAGXVC.tar	294	BBAAAAAAGXVC	DA0169JC	
328	1	urn:nbn:sk:cda-BBAAAAAAGXVD	33 195 819 008	/archfs/sync/2734/2289/urn_nbn_sk_cda-BBAAAAAAGXVD.tar	294	BBAAAAAAGXVD	DA0169JC	
329	1	urn:nbn:sk:cda-BBAAAAAAGXYE	33 168 556 032	/archfs/sync/2828/2289/urn_nbn_sk_cda-BBAAAAAAGXYE.tar	294	BBAAAAAAGXYE	DA0169JC	DA0173JC
333	1	urn:nbn:sk:cda-BBAAAAAAGXBK	32 981 909 504	/archfs/sync/2121/2289/urn_nbn_sk_cda-BBAAAAAAGXBK.tar	294	BBAAAAAAGXBK	DA0173JC	
334	1	urn:nbn:sk:cda-BBAAAAAAGXKT	33 161 216 000	/archfs/sync/2409/2289/urn_nbn_sk_cda-BBAAAAAAGXKT.tar	294	BBAAAAAAGXKT	DA0173JC	
335	1	urn:nbn:sk:cda-BBAAAAAAGXLB	33 158 070 272	/archfs/sync/2422/2289/urn_nbn_sk_cda-BBAAAAAAGXLB.tar	294	BBAAAAAAGXLB	DA0173JC	
336	1	urn:nbn:sk:cda-BBAAAAAAGXLG	33 166 458 880	/archfs/sync/2427/2289/urn_nbn_sk_cda-BBAAAAAAGXLG.tar	294	BBAAAAAAGXLG	DA0173JC	
337	1	urn:nbn:sk:cda-BBAAAAAAGXLK	33 184 284 672	/archfs/sync/2431/2289/urn_nbn_sk_cda-BBAAAAAAGXLK.tar	294	BBAAAAAAGXLK	DA0173JC	
338	1	urn:nbn:sk:cda-BBAAAAAAGXLR	33 162 264 576	/archfs/sync/2438/2289/urn_nbn_sk_cda-BBAAAAAAGXLR.tar	294	BBAAAAAAGXLR	DA0173JC	
339	1	urn:nbn:sk:cda-BBAAAAAAGXLU	33 159 118 848	/archfs/sync/2441/2289/urn_nbn_sk_cda-BBAAAAAAGXLU.tar	294	BBAAAAAAGXLU	DA0173JC	
340	1	urn:nbn:sk:cda-BBAAAAAAHTDT	33 185 333 248	/archfs/sync/2192/2316/urn_nbn_sk_cda-BBAAAAAAHTDT.tar	294	BBAAAAAAHTDT	DA0173JC	
964			31 440 783 239 168			pasok	195	
1907	2	urn:nbn:sk:cda-ABAAAAAAE5XH	33 319 321 600	/archfs/local/2800/2192/urn_nbn_sk_cda-ABAAAAAAE5XH.tar	294	ABAAAAAAE5XH	DA2038JC	
1908	2	urn:nbn:sk:cda-ABAAAAAAE54R	33 166 510 080	/archfs/local/1694/2192/urn_nbn_sk_cda-ABAAAAAAE54R.tar	294	ABAAAAAAE54R	DA2039JC	
1909	2	urn:nbn:sk:cda-ABAAAAAAE5XL	33 008 691 200	/archfs/local/2804/2192/urn_nbn_sk_cda-ABAAAAAAE5XL.tar	294	ABAAAAAAE5XL	DA2039JC	
1923	2	urn:nbn:sk:cda-ABAAAAAARUWF	33 254 539 264	/archfs/local/2767/2627/urn_nbn_sk_cda-ABAAAAAARUWF.tar	292	ABAAAAAARUWF	DA6459JC	
1924	2	urn:nbn:sk:cda-ABAAAAAARVT4	33 836 498 944	/archfs/local/2656/2628/urn_nbn_sk_cda-ABAAAAAARVT4.tar	298	ABAAAAAARVT4	DA6459JC	
1925			25 818 923 608 064			pasok	226	

Obr. č. 1: Ukážka vytvoreného plánu diseminácie

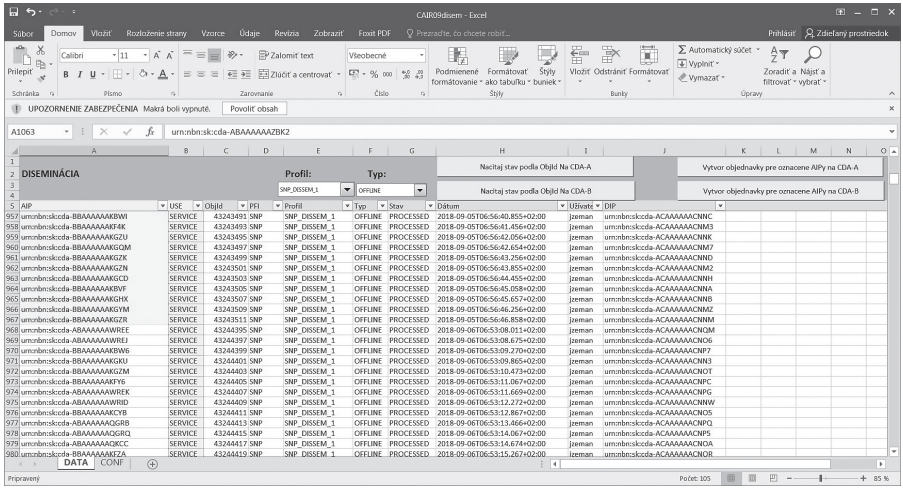
2. Fáza optimalizovaného čítania súborov z pások

Je zabezpečená zadaním príkazu dsmrecall, kde treba špecifikovať ako parameter meno súboru obsahujúceho zoznam požadovaných uložených súborov aj s celou cestou (adresárovou štruktúrou).

30 TiB dát dokáže systém načítať do diskového poľa za približne 11 hodín. Uvedenú fázu zabezpečuje archívny systém IBM Spectrum Protect automatizovane a bez zásahu administrátora.

3. Fáza zadania objednávok pre diseminácii načítaných balíkov.

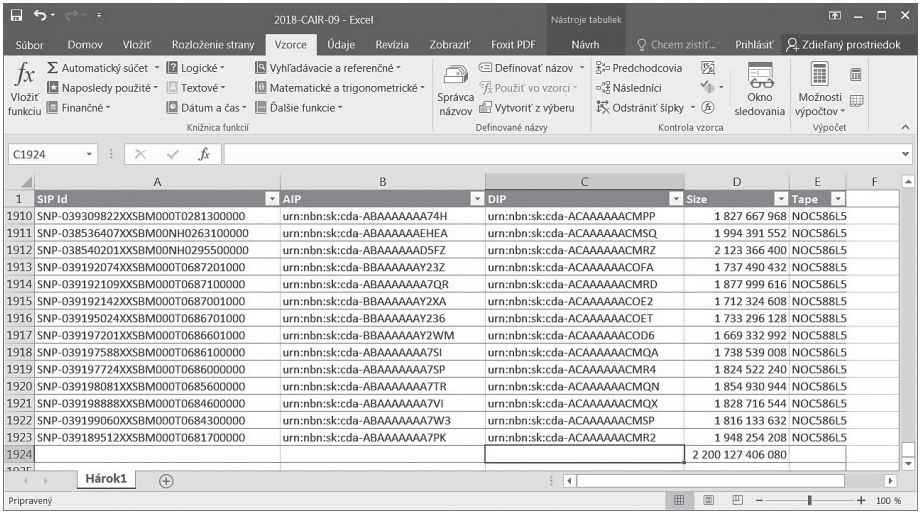
Je zabezpečená pomocou excelového formulára dissemexcel, ktorý pomocou makier napísaných v jazyku VBA zabezpečuje komunikáciu s aplikačným serverom cez CLI rozhranie s použitím užívateľského prístupového certifikátu. Najprv je potrebné si nastaviť pod ktorým PFI užívateľským profilom sa bude diseminácia vykonávať. Následne vyberieme riadky s názvami AIP balíkov. Pre vybrané riadky s uvedením mena AIP balíka postupne vygeneruje požiadavku na diseminácii.



Obr. č. 2: Ukážka objednávkového formulára dissemexcel

4. Fáza povolenia migračného procesu

Je zabezpečená vymazaním podadresára disem_lock vytvoreného vo fáze A. Následne pravidelne spúšťaný migračný proces vyhodnotí stav a zabezpečí vymazanie duplicitných súborov z diskového filesystému, ak už tieto existujú zapísané na páskových médiách.



	SIP Id	AIP	DIP	Size	Tape
1910	SNP-039309822XXSBM000T0281300000	urn:nbn:sk:cda-ABAAAAAAA74H	urn:nbn:sk:cda-ACAAAAAACMPP	1 827 667 968	NOC586L5
1911	SNP-038536407XXSBM00NH0263100000	urn:nbn:sk:cda-ABAAAAAAEHEA	urn:nbn:sk:cda-ACAAAAAACMSQ	1 994 391 552	NOC586L5
1912	SNP-038540201XXSBM00NH0295500000	urn:nbn:sk:cda-ABAAAAAAD5FZ	urn:nbn:sk:cda-ACAAAAAACMRZ	2 123 366 400	NOC586L5
1913	SNP-039192074XXSBM000T0687201000	urn:nbn:sk:cda-BBAAAAAAY23Z	urn:nbn:sk:cda-ACAAAAAACOFA	1 737 490 432	NOC588L5
1914	SNP-039192109XXSBM000T0687100000	urn:nbn:sk:cda-ABAAAAAAA7QR	urn:nbn:sk:cda-ACAAAAAACMRD	1 877 999 616	NOC586L5
1915	SNP-039192142XXSBM000T0687001000	urn:nbn:sk:cda-BBAAAAAAY2XA	urn:nbn:sk:cda-ACAAAAAACOE2	1 712 324 608	NOC588L5
1916	SNP-039195024XXSBM000T0686701000	urn:nbn:sk:cda-BBAAAAAAY236	urn:nbn:sk:cda-ACAAAAAACOET	1 733 296 128	NOC588L5
1917	SNP-039197201XXSBM000T0686601000	urn:nbn:sk:cda-BBAAAAAAY2WM	urn:nbn:sk:cda-ACAAAAAACOD6	1 669 332 992	NOC588L5
1918	SNP-039197588XXSBM000T0686100000	urn:nbn:sk:cda-ABAAAAAAA7SI	urn:nbn:sk:cda-ACAAAAAACMQA	1 738 539 008	NOC586L5
1919	SNP-039197724XXSBM000T0686000000	urn:nbn:sk:cda-ABAAAAAAA7SP	urn:nbn:sk:cda-ACAAAAAACMR4	1 824 522 240	NOC586L5
1920	SNP-039198081XXSBM000T0685600000	urn:nbn:sk:cda-ABAAAAAAA7TR	urn:nbn:sk:cda-ACAAAAAACMQN	1 854 930 944	NOC586L5
1921	SNP-039198888XXSBM000T0684600000	urn:nbn:sk:cda-ABAAAAAAA7VI	urn:nbn:sk:cda-ACAAAAAACMQX	1 828 716 544	NOC586L5
1922	SNP-039199060XXSBM000T0684300000	urn:nbn:sk:cda-ABAAAAAAA7W3	urn:nbn:sk:cda-ACAAAAAACMSP	1 816 133 632	NOC586L5
1923	SNP-039189512XXSBM000T0681700000	urn:nbn:sk:cda-ABAAAAAAA7PK	urn:nbn:sk:cda-ACAAAAAACMR2	1 948 254 208	NOC586L5
1924				2 200 127 406 080	

Obr. č. 3: Ukážka výstupnej tabuľky s evidenciou DIP balíkov pre PFI

Záver

Uvedený postup optimalizovanej diseminácie zabezpečuje minimalizáciu prístupov na magnetické médiá a zároveň dosahuje maximálnu možnú rýchlosť kopírovania údajov danú použitou technológiou. Teda sme schopní v ideálnych podmienkach sprístupniť jeden cyklus (33TiB) za jeden deň. Výstupom sú teda diseminované balíky DIP uložené buď na magnetických páskach typu LTO, alebo sú nakopírované do diskového poľa webdavu pre online prístup PFI.

Použité skratky:

AIP – Archival Information Package (Archívny informačný balík).
DIP – Dissemination Information Package (Diseminačný informačný balík).
PFI – Pamäťová a fondová inštitúcia.
LTO – Linear Tape Open (Otvorený štandard pre páskové médium)
LTFS – Linear Tape File System (Súborový systém pre páskové médium)
CLI – Command line interface (Príkazový interpret)
VBA – Visual Basic for Application (Aplikačný programovací jazyk Visual Basic)
CDA – Centrálny dátový archív
TiB – Tebibyte = 1024*1024*1024*1024*Byte (binárny výpočet)
IBM – Ochranná známka výrobcu HW a SW

Zoznam autorov

Milan Rakús, Juraj Strnisko
Univerzitná knižnica v Bratislave, SR

Zuzana Kvašová
Národní knihovna České republiky, ČR

Piotr Pałka, Tomasz Traczyk
Warsaw University of Technology, Poland

Miklós Lendvay
National Széchényi Library, Hungary

Jiří Bernas
Národní archív České republiky, ČR

Petr Kukač
Národní knihovna České republiky, ČR

Zdeněk Vašek, Petr Cajthaml, Eliška Pavlásková
Ústav dejín Univerzity Karlovy a Archív Univerzity Karlovy, ČR

Zoltán Lux
National Archives of Hungary, Hungary

Stanislav Dzúrik
IBM Slovakia, SR

Márton Németh, László Drótos
National Széchényi Library, Hungary

Andrej Bizík
Univerzitná knižnica v Bratislave, SR

Jaroslav Zeman
Univerzitná knižnica v Bratislave, SR

CDA 2018

Trvalá udržateľnosť a perspektívy ďalšieho rozvoja LTP archívov

Konferencia sa konala v rámci podujatia Týždeň vedy a techniky.

Vydala Univerzitná knižnica v Bratislave
Prvé vydanie. Počet strán 136.

Sadzba: DOLIS GOEN, s.r.o., Bratislava
Tlač: DOLIS GOEN, s.r.o., Bratislava

ISBN 978–80–89303–67–0
ISSN 2453-9309