UNIVERZITNÁ KNIŽNICA V BRATISLAVE

CDA 2017

Výmena skúseností z prevádzky a budovania LTP archívov

Zborník príspevkov z 2. medzinárodnej konferencie o dlhodobej archivácii

univerzitná knižnica
v bratislave

Bratislava, 2017

UNIVERZITNÁ KNIŽNICA V BRATISLAVE

CDA 2017

Výmena skúseností z prevádzky a budovania LTP archívov

Zborník príspevkov z 2. medzinárodnej konferencie o dlhodobej archivácii

univerzitná knižnica
v bratislave

Bratislava, 2017

Zostavil
Ing. Juraj Strnisko

Autori príspevkov
Milan Rakús
Zuzana Kvašová
Monika Péková
Darek Paradowski
Miklós Lendvay
Martin Lhoták
Peter Selecký
Ladislav Cubr
Szabolcs Dancs
Juraj Strnisko
Zdeněk Vašek
Jaroslav Kamenský, Ľubomír Hribík
Roman Král

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

CIP SR

CDA 2017 : Výmena skúseností z prevádzky a budovania LTP archívov : zborník príspevkov z 2. medzinárodnej konferencie o dlhodobej archivácii : Bratislava, 9. 11. 2017 / zost. Juraj Strnisko ; obálka a graf. návrh DOLIS, s. r. o. – 1. vyd. – Bratislava : Univerzitná knižnica v Bratislave, 2017

LTP archívy. Centrálny dátový archív. Dlhodobé dôveryhodné digitálne úložisko. Prevádzka LTP archívov. Štandardizácia. Validátory. Tvorba SIP balíkov. PREFORMA.

ISBN 978-80-89303-58-8
ISSN 2453-9406

Obsah

Úvod

Prevádzka a rozvoj systémov na spoľahlivú dlhodobú archiváciu digitálneho obsahu patrí k aktuálnym úlohám pamäťových inštitúcií. Je to logické pokračovanie tradičných procesov tvorby, ochrany a sprístupňovania hmotných informačných prameňov a artefaktov, inšpiruje sa nimi a zároveň posúva hranice možností poskytovania služieb pre digitalizovanú spoločnosť. Presun ťažiska pozornosti z oblasti digitalizácie na oblasť správy digitálnych objektov je prirodzený a naliehavý aj z dôvodu masívneho nástupu pôvodných digitálnych prameňov. Sprevádza ho množstvo nových teoretických a metodologických a už aj praktických problémov, ktoré sa objavujú vďaka prvým prevádzkovým skúsenostiam so systémami dlhodobej archivácie. Procesy produkcie a ochrany digitálnych objektov nie sú izolované, prirodzene na seba nadväzujú a navzájom sa ovplyvňujú.

Rozhodujúci význam pre trvalo udržateľný rozvoj systémov LTP má štandardizácia, počnúc makroštruktúrou modelu OAIS až po mikroštruktúru metaúdajov vrátane zachytenia technologických detailov digitalizácie a manipulácie s údajmi. Dôveryhodná dlhodobá digitálna archivácia si vyžaduje čoraz hlbší pohľad na samotný predmet ochrany – digitálne objekty, ich formálne a obsahové vlastnosti a tiež na ich potenciál z hľadiska dlhodobého prežitia. Rôznorodosť formátov, štruktúr a množstvo archivovaných údajov kladie zvýšené nároky pri tvorbe archivačných balíkov, ktorá zahŕňa verifikáciu, obohacovanie a často aj konverziu a agregáciu údajov do prijateľnej formy. Osobitnou kapitolou LTP procesov je manipulácia a logistika narábania s údajmi. Zahŕňa problematiku pamäťových médií, prenosu údajov a ich organizácie, kontroly a spôsobu uloženia ako aj pracovných postupov spracovania. Práve tu sú neoceniteľné skúsenosti z praxe.

Univerzitná knižnica v Bratislave prijala na konci roku 2011 výzvu Operačného programu Informatizácia spoločnosti a v rokoch 2012 – 2014 realizovala národný projekt Centrálny dátový archív (CDA) na dlhodobé uchovávanie kultúrneho obsahu. Analytická príprava, technologická realizácia a implementácia proprietárneho LTP systému vyústila už koncom roku 2012 do vkladu prvého archívneho balíka, dnes, na sklonku roku 2017 má za sebou CDA 3 roky reálnej, náročnej, ale aj úspešnej prevádzky. Obdobné zámery a úspechy zaznamenávame aj v susedných krajinách, v Českej republike, Maďarsku či Poľsku.

Cieľom organizátorov konferencie CDA´2017 je prispieť v medzinárodnom kontexte k úrovni poznania v danej oblasti formou prezentácií a diskusií a výmeny praktických a teoretických poznatkov a názorov. Pozornosť, ktorú touto formou venujeme témam dlhodobej ochrany „digitálnych“ znalostí je zároveň príspevkom k napĺňaniu ambície Univerzitnej knižnice v Bratislave v oblasti vedeckej a výskumnej činnosti a pokračovaním systematickej a dlhodobej spolupráce zainteresovaných expertov a inštitúcií.

Druhý ročník medzinárodnej konferencie CDA 2017: Výmena skúseností z prevádzky a budovania LTP archívov sa uskutočnila dňa 9. 11. 2017 v Univerzitnej knižnici v Bratislave.Príspevky sú v zborníku zoradené v poradí podľa programu konferencie.

Konferencia s medzinárodnou účasťou sa konala pri príležitosti Týždňa vedy a techniky na Slovensku 2017. Cieľom týždňa vedy a techniky na Slovensku 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 pomáhajú riešiť globálne problémy a výzvy. Sme radi, že Univerzitná knižnica v Bratislave pri tejto príležitosti prispela významnou mierou – zorganizovaním medzinárodnej konferencie za účasti významných expertov k téme dlhodobého uchovávania digitálneho obsahu, ktorá sa stala stabilnou platformou na výmenu skúseností v tejto oblasti pre pamäťové inštitúcie nielen na Slovensku, ale aj v okolitých krajinách.

V mene organizátorov konferencie

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

Tri roky prevádzky Centrálneho dátového archívu

Milan Rakús, Univerzitná knižnica v Bratislave

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 tri roky prevádzky (2015 – 2017). Príspevok je zameraný na opis procesov spojených s fungovaním LTP archívu, na skúsenosti s jeho prevádzkou a na možnosti riešenia niektorých problémov, ktoré so sebou prináša dlhodobá ochrana digitálnych dát.

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 zastú-

piť 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.

Základným predpokladom finančnej udržateľnosti výsledkov realizácie projektu CDA OPIS PO2 bolo alokovanie dostatočného množstva finančných prostriedkov na obdobie rokov 2015 – 2020 zo štátneho rozpočtu. V súčasnom období je pre CDA uzavretá Zmluva o poskytovaní servisných služieb (SLA), Čiastková zmluva na poskytovanie služieb podpory NON IKT CDA a sú zabezpečené finančné prostriedky na cyklickú obnovu technickej a technologickej infraštruktúry ako aj finančné prostriedky na ostatné nevyhnutné náklady na prevádzku a personál, a to na celé obdobie rokov 2015 – 2020.

CDA má v súčasnom období za sebou takmer tri roky prevádzky (2015 – 2017).

2 Dosiahnuté výsledky a plnenie ukazovateľov projektu

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

2.1 Určené spoločenstvo CDA

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.

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)

2.2 Štatistika vkladov do CDA k 30. 9. 2017

Štatistika vkladov do CDA k 30. 9. 2017 je uvedená na Obr. č. 1.

IDT	Pamäťová a fondová inštitúcia (PFI)	Počet vložených SIP balíkov v CDA	Objem v TB
SNK	Slovenská národná knižnica	90916	32,36
SNA	Slovenský národný archív	1265253	771,41
SNG	Slovenská národná galéria	10184	14,81
SNP	Múzeum Slovenského národného povstania	144463	2679,344
PUR	Pamiatkový úrad Slovenskej republiky	0	0
SFU	Slovenský filmový ústav	996	448,584
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	16537	0,2623
GEO	Štátny geologický ústav Dionýza Štúra	15324	1,325
TSK	Trnavský samosprávny kraj	0	0
NSK	Nitriansky samosprávny kraj	0	0
UKB	Odbor digitalizácie UKB	1752	42,272
SKN	Slovenská knižnica pre nevidiacich Mateja Hrebendu v Levoči	0	0
	Spolu:	**1545425**	**3990,3673**

Obr. č. 1: Štatistika vkladov do CDA k 30. 9. 2017

Kapacita CDA je naprojektovaná na 25 PB dát.

2.3 Štatistika výberov z CDA k 30. 9. 2017

Štatistika výberov z CDA k 30. 9. 2017 je uvedená na Obr. č. 2.

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	48320	818,3001
PUR	Pamiatkový úrad Slovenskej republiky	0	0
SFU	Slovenský filmový ústav	11	0,796
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	7	0,0006
SKN	Slovenská knižnica pre nevidiacich Mateja Hrebendu v Levoči	0	0
	Spolu:	**48360**	**819,1457**

Obr. č. 2: Štatistika výberov z CDA k 30. 9. 2017

2.4 Plnenie ukazovateľov projektu

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.

3 Centrálny dátový archív ako LTP archív

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].

Ako sa s jednotlivými požiadavkami na LTP archív vyrovnáva CDA UKB je uvedené nižšie.

3.1 Finančné zabezpečenie CDA

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=3080320&l=sk) je podpísaná do 8. 9. 2018. Plánované rozpočty na jednotlivé roky obdobia udržateľnosti sa darí plniť a sú dostatočné.

3.2 Obnova IKT CDA

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

3.3 Organizačné a personálne zabezpečenie CDA

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.

3.4 Formátová ochrana

Formátová ochrana patrí k permanentným procesom dlhodobého uchovávania zvereného obsahu, ktorému venuje CDA mimoriadnu a systematickú pozornosť. Téme bola venovaná aj 1. medzinárodná konferencia CDA 2016: Formátové výzvy LTP (http://cda.kultury.sk/sk/Konferencia_CDA_2016).

CDA v maximálnej možnej miere uprednostňuje otvorené formáty súborov, vrátane kontajnerových. Usiluje sa, aby počet formátov súborov v ktorých sú uložené súbory v archíve bol primeraný poslaniu archívu, a pokiaľ sa dá, aby bol minimálne možný. Orientuje sa na také formáty súborov, ktoré majú alebo budú mať k dispozícii kvalitné validátory. CDA sa snaží, zatiaľ úspešne, eliminovať proprietárne formáty Určeného spoločenstva, ktoré nepovažuje za perspektívne a pri vklade by nemali byť súbory v tomto formáte komplexne testované a potom dlhodobo udržiavané. Odporúča Určenému spoločenstvu vhodnejšie akceptované alebo v blízkej budúcnosti akceptovateľné formáty.

Formáty, ktoré akceptuje CDA k 30. 9. 2017 sú uvedené na Obr. č. 3.

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/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/142	audio/x-wav	DROID	JHOVE	
fmt/156	image/tiff	DROID	JHOVE	
fmt/193	application/octet-stream	DROID	JHOVE (modul BYTESTREAM)	DPX 1.0
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. 2017

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

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 1.16 alebo novší / pngcheck	
fmt/12	image/png	DROID	JHOVE 1.16 alebo novší / pngcheck	
fmt/13	image/png	DROID	JHOVE 1.16 alebo novší / pngcheck	
fmt/14	application/pdf	DROID	JHOVE	PDF 1.0
fmt/15	application/pdf	DROID	JHOVE	PDF 1.1
fmt/16	application/pdf	DROID	JHOVE	PDF 1.2
fmt/95	application/pdf	DROID	veraPDF	PDF/A (1a)
fmt/276	application/pdf	DROID	JHOVE 1.16 alebo novší	PDF 1.7 (podpora validátora je čiastočná)
fmt/289	application/warc	DROID	JHOVE 1.16 alebo novší / warctools	
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. 4: Formáty, ktoré bude CDA výhľadovo akceptovať v budúcnosti

CDA sa rozhodol akceptovať dva kontajnerové formáty pre audiovizuálne dokumenty. PIUD = fmt/5 a PUID = fmt/569. PUID (Persistent Unique Identifier) je unikátny identifikátor formátu súboru registrovaný službou PRONOM (http://www.nationalarchives.gov.uk/PRONOM/Default.aspx). Služba PRONOM, ktorá sa v rámci CDA využíva, poskytuje na identifikáciu obsahu súborov nástroj DROID (Digital Record Object Identification).

K formátovej rozmanitosti. Pripomínam, 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 a 1. 10. 2017.

TOP LEVEL TYPE	Stav k 27. 9. 2016	Stav k 1. 10. 2017	Rozdiel
application	1190	1253	63
audio	144	148	4
font	0	6	6
example	1	1	0
image	56	56	0
message	21	21	0
model	22	23	1
multipart	15	17	2
text	72	73	1
video	78	79	1
Spolu:	**1599**	**1677**	**78**

Obr. č. 5: Počty mime type evidované v registroch IANA

Za rok (27. 9. 2016 – 1. 10. 2017) zaevidovala IANA 78 nových formátov. Poznamenávam, že na najvyššej úrovni identifikácie media type bolo zaregistrované dokonca nové meno (font).

Z pohľadu CDA sa nepodarilo doriešiť problém, ako sa vyrovnať s proprietárnymi formátmi, ktoré CDA nepovažuje za perspektívne a pri vklade by nemali byť súbory v tomto formáte komplexne testované a potom dlhodobo udržiavané. Otázka vytvorenia nového CDA proprietárneho kontajnerového formátu, v ktorom budú uložené, z pohľadu CDA súbory v neperspektívnych formátoch, nie je ešte uzavretá.

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) (Pozri Kap. č. 3.10 Medzinárodná spolupráca).

V prvej polovici roku 2017 boli z radov zamestnancov CDA UKB určení dvaja odborníci, ktorí sa systematicky zaoberajú aj problematikou formátov a formátovej stratégie CDA.

3.5 Bitová ochrana (kontrola integrity)

Bitová ochrana (kontrola integrity) patrí medzi permanentné procesy dlhodobého uchovávania zvereného obsahu. Bitová ochrana (kntrola integrity) dát je po troch rokoch prevádzky CDA mimoriadne aktuálna.

V súčasnom období zabezpečuje bitovú ochranu dát na najnižšej úrovni samotný HW. V CDA rozumieme pod bitovou ochranou (kontrolou integrity) dát proces, v rámci ktorého sa archívny balík (AIP) diseminuje, rozbalí, preveria sa kontrolné súčty a vykoná sa antivírová kontrola súborov. Výsledky sa zapíšu do Katalógu CDA. Následne sa zistené nedostatky odstránia.

V CDA-B sa bitová ochrana (kontrola integrity) vykonáva od 23. 8. 2017. K 30. 9. 2017 bolo skontrolovaných 4019 AIP (31,87986 TiB). Kontrolujú sa AIP balíky, ktoré boli naposledy kontrolované alebo vložené do archívu viac ako 1170 dní dozadu od aktuálneho dátumu. Bitová ochrana dát archivovaných v CDA-A a CDA-C je v procese riešenia.

3.6 Synchronizácia lokalít CDA

Z hľadiska ochrany uchovávaných dát je synchronizácia lokalít CDA-A a CDA-B veľmi dôležitý proces, ktorý je dôsledne uplatňovaný od spustenia CDA do prevádzky. Obidve lokality by mali byť identické. Z prevádzkových dôvodov sa stávajú identickými s určitým časovým oneskorením, ktoré je podmienené možnosťou prepravy magnetických pások so synchronizačnými dátami medzi lokalitami.

3.7 Certifikácia CDA podľa normy SMIB

CDA UKB bol v období rokov 2014 – 2016 certifikovaný v súlade s normou „STN ISO/IEC 27001:2013 (SMIB) [6]. V roku 2014 sa uskutočnila samotná certifikácia v rokoch 2015 a 2016 sa realizoval dozorný audit. Certifikácia na ďalšie obdobie podľa uvedenej normy sa v CDA uskutoční do konca roku 2017.

3.8 Audit a certifikácia CDA ako LTP archívu

Existuje niekoľko metód auditu certifikácie LTP archívov (procesov preukazovania skutočnosti, že archív je dlhodobý dôveryhodný archív). Niektoré z nich, používané v európskom priestore, sú uvedené v [5].

CDA UKB má rozpracované a pripravené podklady pre audit a certifikáciu LTP archívu CDA podľa týchto noriem a metodík:

- STN ISO 16363:2014 [7] (pripravené v rokoch 2013 – 2014),
- DRAMBORA (Digital Repository Audit Method Based on Risk Assessment) (http://www.repositoryaudit.eu/) (pripravené v roku 2015),
- DSA (Data Seal of Approval) Pečať kvality digitálneho repozitára (pripravené v roku 2016).

Problém auditu certifikácie archívu ako LTP archívu spočíva v nedostatku certifikačných autorít a vo finančnej náročnosti procesu. Napr. na audit a certifikáciu LTP archívu v súlade s normou STN ISO 16363:2014 [7] existuje vo svete v súčasnom období len jedna autorita. Ako najschodnejšia cesta auditu a certifikácie LTP archívu sa javí audit a certifikácia podľa DSA (Data Seal of Approval) Pečať kvality digitálneho repozitára. Na certifikáciu LTP archívu podľa tejto metódy existuje viac certifikačných autorít.

Audit a certifikácia LTP archívov je oblasť, kde sa dá na medzinárodnej úrovni veľmi tesne a účinne spolupracovať.

3.9 Veda a výskum

Problematika LTP archívov vytvára široký priestor pre aktivity v oblasti vedy a výskumu na inštitucionálnej, národnej aj medzinárodnej úrovni.

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“.

Progresívny charakter mala aj 1. medzinárodná konferencia CDA 2016: Formátové výzvy LTP, ktorá sa uskutočnila 10. 11. 2016 v UKB a má aj 2. medzinárodná konferencia CDA 2017: Výmena skúseností z prevádzky a budovania LTP archívov, ktorá sa koná dnes, 9. 11. 2017, takisto v UKB.

3.10 Medzinárodná spolupráca

CDA UKB je externým partnerom (http://www.preforma-project.eu/external-partners.html) medzinárodného projektu PREFORMA (http://www.preforma-project.eu/project.html).

V rámci projektu PREFORMA sú vyvíjané open source validátory:

- VeraPDF (VeraPDF) na validáciu súborov PDF/A
- DPFmanager (EasyInnova) na validáciu súborov TIFF
- MediaConch (MediaArea) na validáciu súborov MKV (FFV1, PCM)

Výsledky projektu PREFORMA majú pre CDA mimoriadny význam vzhľadom na to, že ich bude perspektívne uplatňovať vo všetkých troch oblastiach validácie formátov súborov. Projekt končí 31. 12. 2017.

Okrem testovania vyvíjaných validátorov sa v rámci riešenia projektu PREFORMA zúčastnili zamestnanci CDA UKB týchto konferencií:

- Berlín, 23. 11. 2016 (http://www.preforma-project.eu/experience-workshop.html)
- Padova, 7. 3. 2017 (http://www.preforma-project.eu/workshop-in-padua.html)
- Tallinn, 20. – 21. 10. 2017 (http://www.preforma-project.eu/final-conference.html)

Medzinárodné kolokvium, Brno 31. 5. – 1. 6. 2016, Knižnice krajín V4 v digitálnom veku, poskytlo rámec na spoluprácu knižníc v rámci 4 základných platforiem.

Už v roku 2016 sme neformálne koncipovali LTP platformu krajín V4. V rámci LTP platformy krajín V4 sa uskutočnilo niekoľko pracovných stretnutí vybratých zástup-

cov 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 sme naplánovali na 10. 11. 2017 opäť v Univerzitnej knižnici v Bratislave. Predpokladáme, že stretnutia sa zúčastnia aj zástupcovia Poľska.

Obsahovou náplňou LTP platformy krajín V4 je spoločné riešenie otázok súvisiacich s problematikou dlhodobej archivácie digitálnych objektov.

3.11 Propagácia CDA LTP archívu

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.

3.12 Spolupráca CDA s Určeným spoločenstvom

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.

4 Záver

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

V CDA UKB sú k dispozícii rozpracované dva strategické materiály: Krátkodobý výhľad rozvoja CDA (do konca obdobia udržateľnosti projektu) a Dlhodobý výhľad rozvoja CDA (po skončení obdobia udržateľnosti projektu).

V rámci Krátkodobého výhľadu rozvoja CDA uvažujeme napr.:

- Zorganizovať medzinárodnú konferenciu CDA 2018: Dlhodobá ochrana dát v LTP archívoch
- 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 DSA (Data Seal of Approval) Pečať kvality digitálneho repozitára

V rámci Dlhodobého výhľadu rozvoja CDA uvažujeme napr.:

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

Na úplný záver chcem poďakovať svojim kolegom Ing. Stanislavovi Lichému a Ing. Alene Špánikovej za pomoc s prípravou vecných podkladov pre tento príspevok.

Použité skratky

AIP – Archival Information Package (Archívny informačný balík)
CDA – Centrálny dátový archív
DIP – Dissemination Information Package (Výberový informačný balík)
HW – Hardvér
IDT – Identifikátor Pamäťovej a fondovej inštitúcie
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 (http://www.opis.gov.sk/)

[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

Three years of operation of the Central Data Archive

Milan Rakús, University Library in Bratislava

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 three years of operation (2015-2017). This paper is aimed at describing the processes involved in the LTP archive, in its experience of operation and in solving some of the problems that the long-term protection of digital data brings with it.

LTP úložiště NK ČR a zkušenosti s jeho provozem

Zuzana Kvašová, Národní knihovna České republiky

Abstrakt
NK ČR provozuje od roku 2013 dlouhodobé úložiště digitálních dokumentů, které vzniklo v rámci projektu Vytvoření Národní digitální knihovny. V souvislosti s potřebami tohoto projektu bylo nastaveno pro přijímání digitalizovaných dokumentů podle standardu NDK, které jsou vytvářeny v digitalizačních linkách partnerských institucí Národní knihovny a Moravské zemské knihovny. Postupně se otevírá i dalším typům dokumentů, nejen z těchto institucí. Příspěvek bude zaměřen na popsání principů fungování současného LTP úložiště, zkušeností s jeho provozem a představení vizí pro budoucí rozvoj.

Úvod

V letech 2009 – 2014 byl ve spolupráci Národní knihovny České republiky a Moravské zemské knihovny realizován projekt Vytvoření národní digitální knihovny (dále projekt NDK)[1], financovaný v rámci priority eCulture dotací ze strukturálních fondů Integrovaného operačního programu EU v kombinaci se spoluúčastí státního rozpočtu ČR. V rámci projektu vznikla v obou knihovnách digitalizační pracoviště a byl vytvořen společný systém pro digitalizaci, zpracování, uložení a zpřístupnění digitalizovaného obsahu. Digitalizováno a zpřístupněno bylo více než 26 milionů stran bohemikálních dokumentů 19. – 21. století. Při takovém množství dokumentů byl kladen velký důraz na zabezpečení dlouhodobého uložení těchto dat, kterému se Národní knihovna ČR začala systematicky věnovat (1).

V rámci projektu proto vzniklo dlouhodobé úložiště (též LTP úložiště) digitálních dat, které následuje principy normy OAIS – ISO norma 14721 (2) definující požadavky systému na zajištění dlouhodobého uložení digitálních dat, tzn., pojmenovává jednotlivé prvky a procesy, které mají vést k ochraně uložených informací nezávisle na sou-

1 Web projektu http://www.ndk.cz/

časném kontextu a využitých technologiích. Pilotní provoz dlouhodobého úložiště byl zahájen v roce 2012, od roku 2013 je úložiště provozováno v Národní knihovně ČR. Data z obou institucí jsou nadále ukládána i po dobu fáze udržitelnosti projektu NDK, která má trvat do roku 2019.

Snahy o dlouhodobou archivaci digitálních dat v Národní knihovně ČR zahrnují komplexní péči o procesy ochrany dat, počínaje výběrem preferovaných standardů a formátů a konče kontrolami a aktualizacemi již existujících uložených dat. Tyto činnosti má na starosti specializovaný Odbor digitálních fondů, který má v kompetenci správu digitálních dat a metodiky pro jejich vznik a návazné fáze jejich životního cyklu.

Provoz LTP úložiště a zajištění dlouhodobé archivace digitálních dokumentů jsou rovněž součástí statutárních činností. Podle zřizovací listiny z roku 2011 Národní knihovna ČR "formuluje strategie a postupy dlouhodobé ochrany elektronických dokumentů a provozuje důvěryhodné digitální úložiště" (3).

Současná politika

LTP úložiště bylo v rámci projektu Národní digitální knihovny koncipováno jako úložiště pro produkci digitalizačních linek partnerských institucí. Na přelomu let 2011 a 2012, kdy vznikaly prováděcí návrhy na vytvoření systému, byla vydána přesná specifikace očekávaných vstupních digitalizačních balíčků (tzv. SIP balíčků). Specifikace zahrnovala návrh struktury digitalizačního balíčku, využitých metadatových standardů a jejich konkrétní atributy a povinnosti, identifikátory, profily archivní a uživatelské kopie v uloženém archivačním balíčku (4). Navíc vznikl návrh tzv. druhotného SIP balíčku, který je využíván pro účely zpřístupnění. Podle těchto návrhů byl LTP systém připraven a předán do provozu. Specifikace byly vydány pro monografie a periodika a v průběhu let je s nimi dále pracováno, jsou upravovány a aktualizovány[2].

Národní knihovna v návaznosti na tento model využívá politiku předem určených specifikací vstupních a archivačních balíčků, které jsou ukládány do LTP úložiště. Tyto specifikace se snaží vystihnout tzv. signifikantní vlastnosti, vhodné pro dlouhodobou ochranu (tedy vlastnosti digitálních dat, které mají být přítomny z důvodu jejich nezbytnosti pro použitelnost, přístupnost a srozumitelnost digitální informace v budoucnosti) (5). Specifikace jsou vydávány na národní úrovni jako standardy pro digitalizaci. Producenti digitálních dat jsou pak vybízeni k tomu, aby výsledná digitální data co nejvíce odpovídala těmto profilům. Národní knihovna si od tohoto přístupu slibuje usnadnění péče o digitální data a navíc vznik poměrně konzistentních dat napříč digitalizacemi v České republice.

2 Standardizace NDK na http://www.ndk.cz/standardy-digitalizace/metadata

Vzhledem k neustálému vývoji je třeba standardy aktualizovat a vydávat nové verze. Specifikace taktéž vznikají i pro nové typy dokumentů (např. pro zvukové dokumenty a elektronické publikace). Využívány jsou jak partnery projektu NDK Národní knihovnou a Moravskou zemskou knihovnou, tak dalšími institucemi v ČR, které digitalizují. Některé instituce je využívají na základě podmínek dotačních programů pro digitalizaci, kde je jejich využívání vynucováno, jiné tak činí dobrovolně. Pro komunikaci s těmito institucemi byl v Národní knihovně vytvořen tzv. Formátový výbor NDK, kde jsou standardy a jejich vývoj projednávány[3].

Výhodou přístupu vydávání přesných specifikací digitálních dat pro ukládání do LTP úložiště je kromě již zmíněné homogennosti dat i v určitém smyslu zjednodušení procesu přijímání dat do úložiště, především díky minimálním nárokům na tzv. „normalizaci" dat, tedy potřebu data upravovat do požadovaných formátů.

Nevýhodou je pak velká náročnost na úpravy specifikací, s vědomím, že každá změna se musí v systému promítnout minimálně změnou validací, při větších změnách pak i úpravou samotného vnitřního formátu LTP úložiště.

Provoz LTP

Jak již bylo uvedeno, LTP úložiště je součástí systému NDK. Je složeno ze tří komponent. Samotné LTP, které zahrnuje LTP SAFE + LTP WF, tedy systém pro uložení a uživatelské rozhraní a systém pro práci s archivačními balíčky. Transformační modul, který řídí vykonávání akcí nad balíčky a modul úložiště – IBM Information Archive, čili systém pro vlastní uložení dat. IBM Information archive je autonomní systém, který ukládá data na pásky.

Data jsou ukládána ve 3 identických kopiích na – 1x online a 2x offline na oddělených lokalitách v Praze a v Brně.

Základní role

Obsahový správce LTP úložiště

Správce obsahu má přístup ke všem uloženým digitálním dokumentům a k objektům souvisejícím s dlouhodobým uchováváním těchto digitálních objektů, což zahrnuje:

- záznamy o importech a exportech balíčků AIP (včetně statistik),
- informace o zpracování balíčků AIP,
- záznamy o kontrolách integrity dat,
- Registr souborových formátů NDK,

3 http://www.ndk.cz/formatovy-vybor-ndk

- Registr metadatových formátů NDK,
- dokumenty plánování uchovávání,
- dokumenty opatření,
- záznamy producentů/dodavatelů dat včetně smluv uzavřených či uzavíraných s těmito dodavateli.

Správce obsahu může provádět export balíčků AIP za účelem vytváření balíčků DIP, deaktivace balíčku AIP, dalšího zpracování balíčku AIP v LTP Workflow (např. při kontrole a opravě metadat) apod.

Technický správce LTP úložiště NDK
Technický správce LTP úložiště NDK zajišťuje technickou funkcionalitu Systému LTP úložiště NDK. Stará se o instalaci, aktualizaci a konfiguraci celého Systému LTP úložiště NDK nebo jeho částí. Má administrátorský přístup do Systému LTP úložiště NDK i do použitých databází a úložišť.

Technický správce LTP úložiště NDK má stejná oprávnění jako Správce obsahu a navíc disponuje administrátorskými právy např. pro správu uživatelů systému, hromadné operace s objekty či správu logových záznamů.

Operátor IA
Operátor IBM information archive vykonává běžné rutinní operace se systémem IBM IBM Information Archive včetně manipulace s páskami, na kterých jsou uložena digitální data uchovávaná LTP úložištěm NDK. Mezi jeho kompetence patří správa záznamů s informacemi o kontrolách integrity dat archivačních balíčků. (6)

Vstup dat do LTP úložiště
Do LTP úložiště v současnosti vstupují archivační balíčky podle Standardu NDK pro tištěné monografické dokumenty a periodika. Většina dokumentů pochází z digitalizační produkce projektu NDK, část dat pochází od externích dodavatelů. Před vstupem do LTP dochází k validaci metadat, k validaci souborových formátů a ke kontrole integrity balíčku na základě kontroly MD5. Ingest probíhá automatizovaně, je řízen tranformačním modulem.

V testovacím provozu je připravován vstup pro elektronické publikace a periodika, který do Národní knihovny odevzdávají vydavatelé na základě dobrovolného odevzdávání.

SAFE LTP

Balíčky, které jsou do LTP ukládany (archivační balíčky, AIP), jsou vytvářeny tak, aby byly samonosné, tedy informace, kterou nesou, byla čitelná nezávisle na konkrétním systému LTP úložiště (7). Odpovídají výše popsaným specifikacím. Archivační balíček je adresář, který obsahuje datové a metadatové soubory. Při uložení se adresář balíčku rozdělí na metadatovou a datovou část a obě se samostatně zazipují. Následně jsou uloženy do úložiště. Při případných změnách a nutných aktualizacích balíčků, které se většinou týkají pouze metadat, je tedy možné aktualizovat pouze metadatový soubor a zvlášť ho verzovat. Datová část archivačního balíčku může zůstat beze změny.

Samotné LTP SAFE je systém pro správu balíčků s uživatelským rozhraním, do kterého jsou vytěžována metadata z archivačních balíčků. Obsahuje vlastní relační databázi (SQL), která umožňuje rychlé operaci nad vytěžovanými údaji a přehled pro uživatele. Do relační databáze se ukládá kopie výše popsané metadatové části (metadatový ZIP) archivačního balíčku.

V rámci subsystému LTP SAFE probíhají automatické i manuální kontroly. Automaticky je kontrolována integrita dat na náhodně vybraných balíčcích v režimu denních kontrol na základě MD5. Manuální kontroly provádí obsahoví správci většinou na základě upozornění z jiných částí systému NDK – ze samotné digitalizace, nebo aplikace pro zpřístupnění.

V rámci systému NDK je držena konzistence dat mezi LTP úložištěm a aplikací pro zpřístupnění (digitální knihovna Kramerius). Čili všechna data, která je třeba opravit z hlediska uživatelského komfortu, jsou opravena již v LTP úložišti a následně v aplikaci pro zpřístupnění.

Systém LTP SAFE obsahuje vlastní editační modul (kopii modulu, který je využíván v digitalizační lince NDK). Ten umožňuje provádět kontroly dokumentů a některé typy oprav archivačních balíčků.

V systému LTP SAFE existuje několik procesů, kterými je možné provádět opravy. Jedná se o:

- Deaktivace – Proces, při kterém dojde k deaktivování archivačního balíčku, data však fyzicky zůstávají uložena na páskách. Dojde ke smazání uživatelských kopií z aplikace pro zpřístupnění (Kramerius). Následně je smazán identifikátor URN:NBN a odkazy z katalogu Aleph na digitalizovaný dokument. Manuálně musí být smazán údaj o digitalizaci v Registru digitalizace.

- Export archivních obrazových dat pro nové zpracování – Proces, kterým je možné opravit metadata v archivačním balíčku, rozdělit balíček na nové intelektutuální entity atd.
- Resken/dosken z LTP – Proces, kterým je možné opravit obrazová data v rámci intelektuální entity, případně i špatně editované údaje z postprocessingu. Není možné měnit strukturu a typ dokumentu a většinu bibliografických údajů z katalogu.
- Aktualizace metadat z Alephu – Tento proces umožňuje aktualizovat bibliografické údaje z katalogu Aleph, pokud nedojde ke změně typu dokumentu. Opravenému dokumentu je vždy přidělen nový identifikátor URN:NBN s vazbou na předchozí identifikátoru. Po vzniku aktualizace balíčku v LTP jsou data aktualizována i v digitální knihovně.
- Oprava dat v EM LTP – Proces opravy dat v EM umožňuje odmazání stran v dokumentu (bez náhrady za nové obrazové soubory), editaci povolených polí v popisných metadatech a opravu UUID na úrovni ročníku a titulu.
- Oprava dat správcem LTP – Umožňuje ve spolupráci s technických správcem LTP opravit větší množství (dávku) dokumentů. Přesná funkcionalita se nastavuje vždy pro každý typ opravy, v závislosti na tom se potom může, nebo nemusí přidělit nový identifikátor URN:NBN, aktualizují se údaje v digitální knihovně atd.
- Hromadná oprava UUID u periodik – Tento proces je upřesňován konfiguračním souborem, který definuje typ opravy UUID. Následně dojde k automatické aktualizaci identifikátoru UUID na úrovni ročníku, nebo titulu periodika. Opravy v digitální knihovně se dělají jak manuálně (smazání dokumentů s chybným UUID), tak automaticky – na základě spuštění procesu exportu „Doplnění K4“.

Archivační balíčky lze tedy v rámci dlouhodobého úložiště aktualizovat, nicméně vždy zůstává na páskách uložena první verze, která je nadále ochraňována na úrovni bit-stream.

Primárním přístupem, jak zabezpečit dlouhodobou archivaci je uchovávat data v takových formátech, aby mohla být zaručena jejich čitelnost a srozumitelnost. Je proto počítáno s prováděním formátových migrací. Do současnosti zatím nebylo využito této funkcionality nad větším množstvím dat.

Současnost a budoucnost LTP úložiště

V současnosti jsou do LTP úložiště ukládány především textové monografické a periodické dokumenty z produkce NDK a některých dalších digitalizací. V menším množství jsou ukládána data z tzv. historických digitalizací, tedy data z projektů, které běže-

ly před rokem 2012, kdy byl vydán současný standard. Ta prochází úpravou ještě před vstupem do úložiště tak, aby co nejvíce odpovídala současným standardům.

Ukládání dalších typů dokumentů brání především omezená harwarová kapacita celého digitalizačního workflow, které LTP úložiště částečně sdílí s digitalizační linkou. Je proto prioritou oddělit LTP úložiště tak, aby digitalizační linka NDK byla jen jedním z mnoha vstupů.

V rámci České republiky je realizována celá řada digitalizačních projektů různých typů dat – od digitalizace rukopisů a starých tisků, po digitalizaci zvukových dokumentů. Všechny tyto dokumenty si zaslouží dlouhodobé uložení.

Zdroje:

1. *NDK: O projektu* [online]. Praha: Národní knihovna ČR, 2015 [cit. 2017-09-22]. Dostupné z: http://www.ndk.cz/o-projektu
2. ČSN ISO 14721. *Systémy pro přenos dat a informací z kosmického prostoru – Otevřený archivační informační systém – Referenční model*. Praha: Úřad pro technickou normalizaci, metrologii a státní zkušebnictví, 2014. 111 s.
3. *Národní knihovna České republiky: Základní dokumenty* [online]. Praha: Národní knihovna České republiky, 2017 [cit. 2017-09-22]. Dostupné z: http://www.nkp.cz/o-knihovne/zakladni-informace/zakladni-dokumenty
4. LODROVÁ, Iveta, ŠVÁSTOVÁ, Pavla, KVASNICA, Jaroslav. *Definice metadatových formátů pro digitalizaci monografických dokumentů (monografií, kartografických dokumentů, hudebnin)* [online]. Verze 1.2. Praha, 2015-07-22. Dostupné z: http://www.ndk.cz/standardy-digitalizace/DMFmonografDok_12.pdf
5. GIARETTA, David, et al. *Significant properties, authenticity, provenance, representation information and OAIS information* [online]. California Digital Library, 2009 [cit. 2017-09-22]. Dostupné z: http://escholarship.org/uc/item/0wf3j9cw
6. CUBR, Ladislav. *Pravidla pro zpřístupňování LTP úložiště NK ČR.* Verze 0.2. Praha: Národní knihovna, 2015. Interní dokument k certifikaci DSA.
7. CUBR, Ladislav. *Plán pro zavedení uchovávání LTP úložiště NDK.* Verze 0.2. Praha: Národní knihovna, 2015. Interní dokument k certifikaci DSA.

Elektronický archív Slovenska ako LTP archív

Mgr. Monika Péková, Ministerstvo vnútra Slovenskej republiky, odbor archívov a registratúr, Križkova 7, 811 04 Bratislava, monika.pekova@minv.sk

Kľúčové slová:
digitalizácia, Elektronický archív Slovenska, elektronický archívny dokument, ochrana archívnych dokumentov, preberanie elektronických archívnych dokumentov, štátne archívy

Abstrakt
Príspevok je zameraný na predstavenie projektu Ministerstva vnútra SR, na ktorý archívna komunita čakala už dlhší čas, a to projekt Elektronický archív Slovenska. Elektronický archív Slovenska predstavuje komunikačný bod pre fyzické osoby a právnické osoby pre všetky agendy podľa zákona NR SR č. 395/2002 Z. z. o archívoch a registratúrach. Zabezpečuje zverejňovanie voľne prístupných archívnych dokumentov v ich digitálnej podobe na webovom sídle a zároveň plní úlohy bezpečného uloženia zdigitalizovaných archívnych dokumentov a elektronických archívnych dokumentov vzniknutých z činnosti pôvodcov registratúry. Príspevok informuje o nastavených procesoch ochrany pri preberaní a spravovaní elektronických archívnych dokumentov.

Hlavným cieľom projektu Elektronický archív Slovenska bolo vybudovanie elektronického archívu s portálom, ktorý je prístupný pre občanov, orgány verejnej moci a ostatné právnické osoby na webovom sídle Elektronické služby Ministerstva vnútra Slovenskej republiky[1] a jeho prepojenie s Ústredným portálom verejnej správy a informačnými systémami verejnej správy iných inštitúcií. Implementácia elektronických služieb umožňuje využívať služby poskytujúce prístup k archívnym dokumentom rýchlejším a efektívnejším spôsobom, a to najmä z pohodlia domova.

Úlohou archívov, v súlade so zákonom NR SR č. 395/2002 Z. z. o archívoch a registratúrach a o doplnení niektorých zákonov v znení neskorších predpisov, je všestranná starostlivosť o archívne dokumenty, ich odborné a vedecké spracovanie a sprístupne-

1 dostupné online na https://portal.minv.sk

nie ako aj využívanie na vedecké a iné účely. Starostlivosť o archívne dokumenty spočíva v ich zhromažďovaní v archívoch, preberaní od ich pôvodcov alebo nadobúdaní od ich vlastníkov, ich riadnej evidencii, bezpečnom uložení a ochrane. Elektronický archív Slovenska predstavuje komunikačné centrum vo vzťahu ku všetkým agendám vyššie uvedeného zákona.

Elektronické služby portálu sa delia na služby aplikačného rozhrania a služby používateľského rozhrania. Služby aplikačného rozhrania slúžia na komunikáciu s inými elektronickými službami informačných systémov orgánov verejnej správy. Služby používateľského rozhrania sú elektronické služby poskytované formou elektronických formulárov pre fyzické osoby a právnické osoby. Elektronický archív Slovenska umožňuje podporu elektronizácie evidencie archívneho dedičstva, je integrovaný na interný informačný systém Ministerstva vnútra SR[2], čím sa podporuje sprístupňovanie archívnych fondov prostredníctvom archívnych pomôcok poskytujúcich komplexný pohľad na pôvodcu archívneho fondu a jeho kompetencie, činnosti a agendy.

Nevyhnutnou podmienkou pre využívanie elektronických služieb Elektronického archívu Slovenska, okrem funkcie prezerania a vyhľadávania archívnych dokumentov, je autentifikácia prostredníctvom aktivovaného elektronického občianskeho preukazu s čipom a aktívne prostredie fyzickej osoby či právnickej osoby.

Základnou víziou architektúry riešenia elektronického archívu je vytvorenie robustného centralizovaného riešenia, postaveného na najmodernejších technológiách. Veľmi dôležitým prínosom je centralizácia, zvýšenie bezpečnosti a trvácnosti uchovávania elektronických archívnych dokumentov preberaných od pôvodcov registratúry vo vyraďovacom aj mimo vyraďovacom konaní, ako aj možnosť uchovávania elektronických kópií pôvodne neelektronických archívnych dokumentov. Riešenie zabezpečuje trvalé uloženie elektronických archívnych dokumentov v správe štátnych archívov zriadených Ministerstvom vnútra Slovenskej republiky, ich autenticitu, hodnovernosť, neporušiteľnosť ich obsahu a čitateľnosť, ako aj ich sprístupnenie a zverejňovanie občanom a odbornej i laickej verejnosti. Informačný systém je umiestnený v dátovom centre Ministerstva vnútra Slovenskej republiky.

Spôsoby napĺňania dátového úložiska Elektronického archívu Slovenska

1. Vkladanie elektronických archívnych dokumentov

Elektronický archív Slovenska poskytuje archivárom funkcionalitu pre riadenie preberacieho konania. Preberaniu elektronických archívnych dokumentov predchádza proces

2 aplikačné programové vybavenie pre evidenciu archívneho dedičstva AFondy,

podania návrhu na vyraďovacie konanie registratúrnych záznamov pôvodcom vrátane zoznamu vecných skupín registratúrnych záznamov navrhnutých na vyradenie, proces posúdenia návrhu a vydania rozhodnutia štátnym archívom. Preberacie konanie iniciuje archivár vydaním výzvy na odovzdanie archívnych dokumentov, ktorá sa viaže k vybranému vyraďovaciemu konaniu. Táto výzva je doručená pôvodcovi elektronicky prostredníctvom elektronickej schránky a zároveň je mu na portáli[3] sprístupnený odkaz na nahratie najprv zoznamu odovzdávaných dokumentov a neskôr aj samotných elektronických archívnych dokumentov po dávkach, ktoré sú priebežne spracované. Ukončenie procesu nahrávania elektronických archívnych dokumentov oznamuje pôvodca stlačením príslušného tlačidla. Po odovzdaní elektronických archívnych dokumentov Elektronický archív Slovenska automaticky vypracováva *Protokol o kontrolách zoznamu odovzdávaných dokumentov* a *Protokol o výsledkoch spracovania odovzdaných AD*. Protokol o kontrolách zoznamu odovzdávaných dokumentov obsahuje informácie, či boli v skutočnosti odovzdané tie elektronické archívne dokumenty, ktoré pôvodca uviedol v zozname vyraďovaných dokumentov, resp. či niektoré elektronické archívne dokumenty neboli nahraté do úložiska omylom viackrát alebo niektoré neboli do úložiska vôbec nahrané. Samotné odovzdávanie elektronických archívnych dokumentov má zatiaľ obmedzenie vo forme maximálnej veľkosti jednej dávky archívnych dokumentov 50 MB. Povolené formáty pre zasielanie do dlhodobého úložiska Elektronického archívu Slovenska sú formáty csv, doc, docx, gif, html, jpg, mp3, mp4, pdf, png, ppt, pptx, tiff, txt, xml, xls, xlsx a ukladajú sa vo formátoch jpeg2000, tiff, pdf/A-1a alebo txt[4].

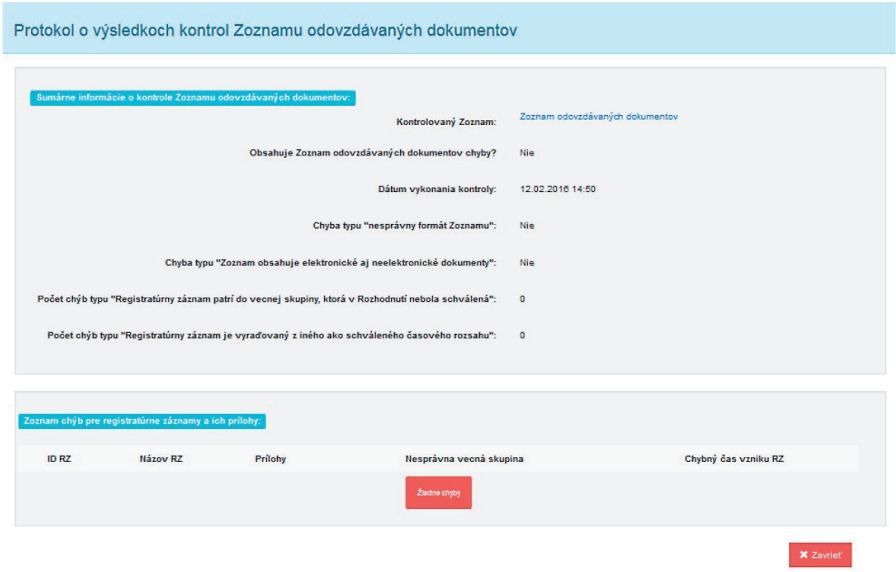
Elektronický archív Slovenska po nahratí súborov elektronických archívnych dokumentov zabezpečuje aj automatické kontroly archívnych dokumentov, ktoré pôvodca nahral. Týmito automatickými kontrolami sú:

1. antivírová kontrola,
2. kontrola integrity súboru,
3. kontrola kompletnosti nahratých súborov oproti zámeru,
4. kontrola formátu súboru,
5. kontrola elektronických podpisov,
6. kontrola nahratia súboru do úložiska.

V prípade chýb v Protokole o výsledkoch spracovania dôjde k prevzatiu iba časti elektronických archívnych dokumentov a pre neprevzatú časť sa vytvorí nové preberacie konanie. Pre prevzaté a uložené archívne dokumenty systém automaticky vytvorí zo-

3 v rámci jeho osobného prostredia

4 § 14a ods. 3 vyhlášky MV SR č. 628/2002 Z. z., ktorou sa vykonávajú niektoré ustanovenia zákona o archívoch a registratúrach a o doplnení niektorých zákonov v znení neskorších predpisov



Obr. 1 Protokol o výsledkoch spracovania odovzdávaných archívnych dokumentov

znam preberaných archívnych dokumentov, ktorý je formou protokolu odoslaný pôvodcovi archívnych dokumentov.

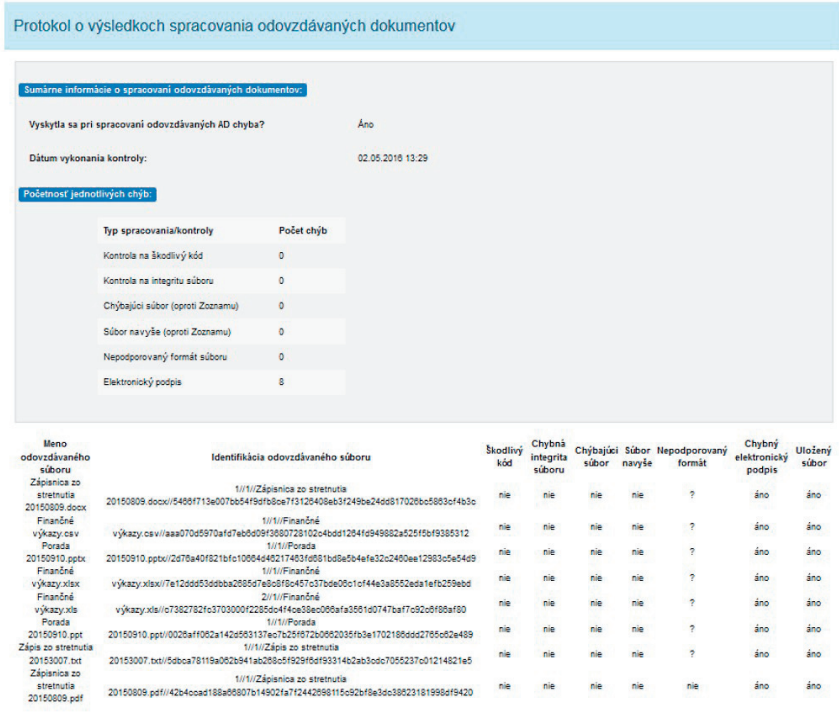
Môžu vzniknúť aj prípady, kedy ani jeden z odovzdaných elektronických archívnych dokumentov neprešiel kontrolami alebo došlo k oznámeniu o ukončení odovzdávania archívnych dokumentov skôr, než samotné dokumenty boli nahrané. V takýchto prípadoch dochádza k vypracovaniu oznámenia o neprijatí archívnych dokumentov.

Ďalším spôsobom ako môžu vlastníci archívnych dokumentov odovzdať elektronické archívne dokumenty do Elektronického archívu je mimo vyraďovacie konanie[5]. Preberacie konanie je podobné procesu preberania archívnych dokumentov ako pri vyraďovacom konaní s tým rozdielom, že vlastník nahrá archívny dokument priamo do úložiska.

2. Vkladanie pôvodne neelektronických archívnych dokumentov

Tretím, a v súčasnosti zatiaľ najrozšírenejším, spôsobom ako možno rozširovať vkladať pôvodne neelektronické archívne dokumenty do úložiska elektronického archívu je funkcia na vkladanie digitalizátov. Táto funkcionalita je prístupná iba štátnym archívom. V rámci tohto procesu archivár uvádza názov konania, čiže dôvod zdigitalizova-

5 ponúknutie archívneho dokumentov na odkúpenie do vlastníctva štátu, dar alebo uloženie do depozitu



Protokol o výsledkoch spracovania odovzdávaných dokumentov

Sumárne informácie o spracovaní odovzdávaných dokumentov:

Vyskytla sa pri spracovaní odovzdávaných AD chyba? Áno

Dátum vykonania kontroly: 02.05.2016 13:29

Početnosť jednotlivých chýb:

Typ spracovania/kontroly	Počet chýb
Kontrola na škodlivý kód	0
Kontrola na integritu súboru	0
Chýbajúci súbor (oproti Zoznamu)	0
Súbor navyše (oproti Zoznamu)	0
Nepodporovaný formát súboru	0
Elektronický podpis	8

Meno odovzdávaného súboru	Identifikácia odovzdávaného súboru	Škodlivý kód	Chybná integrita súboru	Chýbajúci súbor	Súbor navyše	Nepodporovaný formát	Chybný elektronický podpis	Uložený súbor
Zápisnica zo stretnutia 20150809.docx	1//1//Zápisnica zo stretnutia 20150809.docx//5466f713e007bb54f9dfb8ce7f3126408eb3f249be24dd817026bc5883cf4b3c	nie	nie	nie	nie	?	áno	áno
Finančné výkazy.csv	1//1//Finančné výkazy.csv//aaa070d5970afd7eb6d09f3680728102c4bdd1264fd949882a525f5bf9385312	nie	nie	nie	nie	?	áno	áno
Porada 20150910.pptx	1//1//Porada 20150910.pptx//2d78a40f821bfc10664d46217483fd681bd8e5b4efe32c2460ee12983c5e54d9	nie	nie	nie	nie	?	áno	áno
Finančné výkazy.xlsx	1//1//Finančné výkazy.xlsx//7e12ddd53ddbba2685d7e0c8f5c457c37bde06c1cf44e3a8552eda1efb259ebd	nie	nie	nie	nie	?	áno	áno
Finančné výkazy.xls	2//1//Finančné výkazy.xls//c7382782fc3703000f2285dc4f4ce38ec066afa3561d0747baf7c92c6f86af80	nie	nie	nie	nie	?	áno	áno
Porada 20150910.ppt	1//1//Porada 20150910.ppt//0026aff062a142d563137ec7b25f672b0662035fb3e1702186ddd2765c82e489	nie	nie	nie	nie	?	áno	áno
Zápis zo stretnutia 20153007.txt	1//1//Zápis zo stretnutia 20153007.txt//5dbca78119a062b941ab268c5f929f5df93314b2ab3ddc7055237c01214821e5	nie	nie	nie	nie	?	áno	áno
Zápisnica zo stretnutia 20150809.pdf	1//1//Zápisnica zo stretnutia 20150809.pdf//42b4cead188a66807b14902fa7f2442698115c92bf8e3dc38623181998df9420	nie	nie	nie	nie	nie	áno	áno

Obr. 2 Oznámenie o neprijatí odovzdávaných archívnych dokumentov

nia príslušných archívnych dokumentov, objem zdigitalizovaných dát[6] a základné metadáta k digitalizátom.

3. Spravovanie a ochrana archívnych dokumentov

Na úseku ochrany archívnych dokumentov poskytuje elektronický archív podmienky pre trvalé uloženie elektronických archívnych dokumentov a zdigitalizovaných neelektronických archívnych dokumentov. Zaručuje zachovanie integrity, čitateľnosti a aktuálnosti formátov uchovávaných archívnych dokumentov. Všetky procesy sú v zhode s Open Archival Information System (OAIS) štandardom. Elektronický archív Slovenska poskytuje

– bitovú ochranu, ktorá zabezpečuje, že používané páskové média, na ktorých je uložená kópia archivovaného obsahu, bude pred uplynutím ich prevádzkovej doby života prepísaná na nové médium. Administrátor označí média, ktoré chce prepísať a úlohu posunie ďalej na realizáciu. Proces bitovej ochrany sa spúšťa pravidelne, a to raz mesačne.

6 ten sa po schválení alokuje pre nahrávanie v úložisku

– formátovú ochranu, ktorá pozostáva z procesu monitorovania, či všetky formáty v archíve sú aktuálne a z procesu, ktorý zabezpečuje transformáciu archívnych dokumentov v končiacich formátoch do nástupnického formátu. Cieľom procesu je identifikovať formáty archívu, ktoré zastarávajú a vypracovať pre ich obnovu transformačný plán. Proces vykonáva administrátor elektronického archívu, ktorého hlavnou úlohou je z dostupných informačných zdrojov identifikovať, ktorý formát zastaral, zvoliť vhodnú transformačnú technológiu, previesť pilotnú transformáciu a napokon pripraviť transformačný plán.

Cieľom vybudovania informačného systému však nie je iba samotné bezpečné uloženie, ale najmä široké využívanie archívnych dokumentov laickou a odbornou verejnosťou. Tieto úlohy zabezpečuje Elektronický archív Slovenska prostredníctvom portálu Elektronické služby Ministerstva vnútra Slovenskej republiky. V časti *Prezeranie obsahu archívu a vyhľadávanie*[7] sa bádatelia môžu dostať aj k voľne prístupným historickým archívnym dokumentom v ich digitálnej podobe. Prezeranie voľne dostupných archívnych dokumentov nie je viazané na použitie elektronického občianskeho preukazu. V súčasnosti tu zverejňujeme vyše 5 000 digitálnych kópií máp z archívneho fondu Hlavný komorskogrófsky úrad v Banskej Štiavnici a pripravujeme zverejnenie ďalších kópií archívnych dokumentov. Priamo na tomto portáli budú zverejnené aj všetky archívne pomôcky v elektronickej forme, ktorých migrácia práve prebieha.

Záverom konštatujeme, že odbor archívov a registratúr s nadšením privítal realizáciu tohto projektu, a to nielen z dôvodu, že štátne archívy nedisponovali integrovaným systémom dlhodobého uchovávania a ochrany digitálneho obsahu. Za prínos tohto projektu, okrem iného, možno označiť

– zavedenie sofistikovaných elektronických služieb pre občanov, odbornú verejnosť a bádateľov, ktoré zjednodušia možnosti prístupu k archívnym dokumentom a skrátia čas potrebný na sprístupnenie informácií,
– skrátenie času potrebného na predloženie pôvodne neelektronických a elektronických archívnych dokumentov,
– optimalizáciu a zrýchlenie interných procesov štátnych archívov,
– centralizáciu systémov štátnych archívov.

Ďalšia systematická digitalizácia a napĺňanie LTP úložiska vrátane presnej identifikácie archívnych dokumentov vyžaduje zanietených a vzdelaných archivárov a dostatok finančných prostriedkov.

7 dostupné online https://portal.minv.sk/wps/portal/domov/isea/EA07_PrezeranieObsahuArchivu

Step towards self-contained store – challenge, concept and implementation of archival subsystem based on new ISO standard 20919:2016

Dariusz Paradowski, National Library of Poland, al. Niepodleglosci 213, 02-086 Warsaw Poland, +48 22 608 26 17, d.paradowski@bn.org.pl

Abstract
The National Library of Poland (NLP) operates comprehensive digital repository system. Recently NLP replaced commercial appliance that served as archival module by new long time preservation subsystem based on open standards. New software was designed in NLP and makes use of emerging open standard of digital magnetic tape structure – LTFS (ISO 20919:2016). The result is efficient, economic, scalable and safe archival storage component that currently stores over half petabyte of data.

Keywords
Digital archival storage, long time preservation, open standard.

1. CHALLENGE

1.1 Change

Starting from 2014 National Library of Poland (NLP) operates comprehensive system – The Repository – for supporting digitization of library materials, processing created digital files, presentation and storage. The Repository has a number of functional modules responsible for multitrack workflow, import from various sources (scanning, digital legal deposit), conversion of data formats and metadata, generating derivatives (jpg, OCR, etc.) presentation and access management, storage and finally archiving.

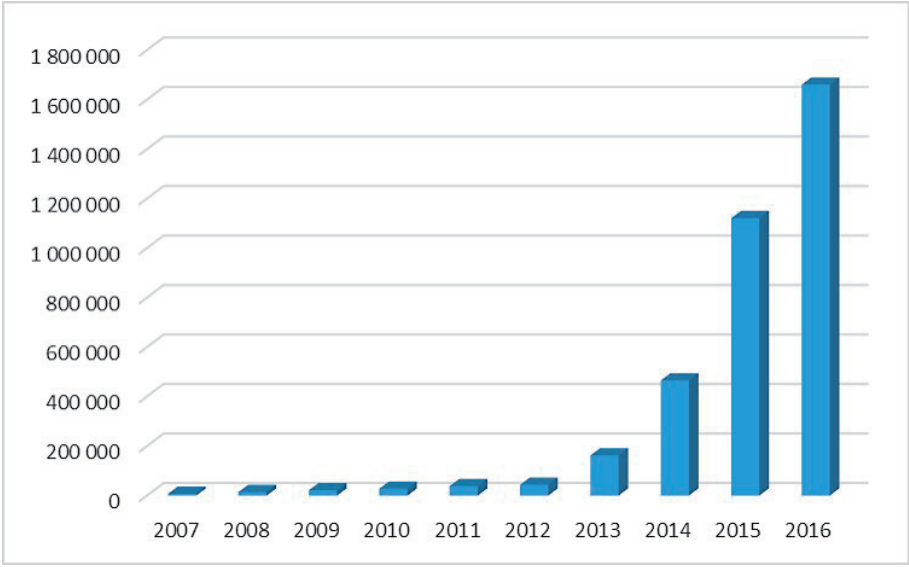


Figure 1. Items available in polona.pl

Recent generation of The Repository immediately after the introduction of the source files to the system put them in an object archive. It is a commercial appliance in the form of licensed software running on a multi-node cluster of dedicated servers connected to the disc array. For some years it was a very useful solution: reliable, efficient and capacity was enough for predictable future.

- it provided a guarantee that extremely important archiving module will be reliable – it was a commercial solution verified in considerable applications,
- cluster equipped with load balancing efficiently acquired and provided files,
- it supplied manageable WORM functions,
- by compression it optimally used disc array capacity, ensuring that the current amount satisfied the needs for a longer period of time.

In recent years digitization efficiency and quality was vastly improved. The result was an exponential growth of data inflow to the Repository.

It this conditions Appliance became inefficient, too expensive and unsafe:

- insufficient performance – necessary increase would require hardware upgrade of the appliance – dedicated servers and arrays and also changes to the architecture of The Repository software
- rapid depletion of capacity – the further operation would require costly expan-

sion of the storage array and also very expensive purchase of licenses for the appliance archive
- increasing energy consumption of the solution
- a black box solution although technically very stable the potential consequences of end of support from producer would be unpredictable – not recommended solution for long time preservation.

1.2 Defined requirements

There were essential needs recognized for new archival system that has to be: safe, efficient, economic and scalable.

These were redefined into more detailed tasks:
- safety
  - open standards
    - the system can not depend on single manufacturer
    - data must be readable outside environment of the archive (without context)
    - storehouse must be easy to transport in case of predictable danger
    - metadata must be human-readable
  - recognized standards
  - durable carrier
  - migration feasible
  - damage to any part of the data can not prevent the reading of data undamaged
- low cost storage
  - low cost of capacity per byte
  - no expensive capacity license
  - low energy consumption
- high, easily scalable performance
  - horizontal scaling possible without rebuilding the infrastructure (just extension) and software
  - vertical scaling possible without software change
- high, easily scalable capacity
  - horizontal scaling possible without rebuilding the infrastructure (just extension) and software
  - vertical scaling possible without software change

2. CONCEPT

2.1 Choice of Carrier

Comparison of separated to composite storage as in case of tape to disc storage. Table 1 shows comparison of the sensitivity of the system disk in a very robust and expensive version of RAID 10 compared with a set of independent carriers comprising two copies of the data. With minor injuries RAID gives greater protection than the worst case for independent media and the same as the best case. It is worth noting that the worst case is relatively unlikely (damage to same data on different tapes) and the best case gives better results than RAID. In particular, the destruction of more than 50% of the media RAID causes a loss of 100% of the data, while the media independent only 25-50%. It is also worth noting that the price of a unit capacity of good quality media is much lower for tape cartridges. It is also important from the point of long time preservation that in case of danger cartridges may be easy removed from tape library and transported which is much more complicated for hard drives. For archival storage separated carriers are usually better than RAID.

Table 1. RAID10 vs independent carriers vulnerability

	Data loss %			
No of carriers destroyed	8 hard disc drives, RAID 10, worst case	8 hard disc drives, RAID 10, best case	2 sets of 4 independent tapes, worst case	2 sets of 4 independent tapes, best case
1	0	0	0	0
2	100	0	25	0
3	100	0	25	0
4	100	0	50	0
5	100	100	50	25
6	100	100	75	50
7	100	100	75	75
8	100	100	100	100

Sequential recording on tape cartridges, which corresponds to scenarios of archival usage is performed with a very high speed. The performance of the system can be easily multiplied by increasing the number of drives and the capacity by increasing number of tapes.

Magnetic tape fulfils a number of requirements:

- It is recognized as carrier for few decades and has verified long shelf life.
- It is easy to transport

- Capacity and speed is scalable by multiplying cartridges and drives.
- Tape has exceptionally low cost per unit of capacity.

It is essential to use an open and yet recognized standard.

Linear Tape Open (LTO) is an open standard supported by many major manufacturers, it also has defined roadmap for development.

- an open standard supported by many major manufacturers
- defined roadmap for development, 2 generations compatible [1]
- new LTO7 on the market at the end of 2015 (6TB, 300 MB/s)
- is supported by automatic tape library already working in NLP

The new standard LTO generation 7 (LTO7) appeared on the market at the end of 2015 and has a capacity of 6TB. This is enough to avoid necessity to purchase another expansion frame for the automatic tape library currently used in NLP in the foreseeable future. It has long time to the end of support and moreover will be readable by next two generations of drives [1].

2.2 Choice of Filesystem

Linear Tape File System (LTFS) meets the requirement of the system to be open, it is supported by several leading manufacturers, developed in the mature form and present on the market for several years and natively supported in LTO.

In the year 2016 LTFS became adopted as standard ISO / IEC [2].

Record in LTFS can be read on another device from another manufacturer, without the need to reconstruct an environment where it was saved. Moreover, basic software solutions – allowing the use LTFS on a single drive are available as open source by many hardware manufactures.

2.3 Choice of metadata container

Versatile and well recognized standard was needed to fulfil requirements. Metadata Encoding and Transmission Standard (METS) has strong theoretical background [3], lively users environment and became standard for metadata exchange and storage has so it was an natural choice for creation of OAIS AIP packages [4].

2.4 Developed rules

A complete library object is understood as: a unique object identifier, all the metadata and structure of the object in the form of (human readable) XML METS and all source content files of the object.
Each entire library object is to be archived on a single carrier. The system is to serve as a disaster recovery solution, so the assumption that there will survive a random subset of cartridges implicates the requirement that each object will be stored on one medium.

Once saved, the object in the archive is never to be changed. Changes in object never results in modification of archived object copy. This prevents backwards error propagation and allows to optimally use tape library. A carrier most of the time is kept offline. In case of addition of new content files to the object, new version of entire object is archived.

Barcode number identifying tape on which each object is stored, is written to database.

All other changes (deletion, order modification etc. of content files or any change in metadata) results in modification of the database which is periodically archived separately (versioned).

Information about the location of the next version (barcode of the cartridge) is to be placed in the database system. It allowed to avoid designing a complex and unreliable predictions of reserved free space on the tape needed to create new versions of files (that way would be also very inefficient considering linear nature of the tape recording).

Problem of archiving metadata, which, in the national library reality are subject to frequent revisions was solved by independent archiving the updated metadata of all objects through saving the entire database of the system in an XML file (with checksums). Likewise objects each database copy is to be kept without adjustments and with versioning instead. Whole database is periodically saved into special database archival package. Thus, if after the disaster, a random set of tape cassettes has been discovered, it is enough to find the latest version of the database which allows to quickly find the latest versions of objects.

3. IMPLEMENTATION

Software design, development and implementation took nine months. Many new problems and challenges appeared and were solved.

Each entire library object is put into a single Tar file for easy handling, checksum verification and tape writing optimization.

On request for object given by REST API The Archive System returns desired version (timestamp) of the object.

During development some innovative algorithms were designed for non-blocking temporary storage use and efficient tape utilization. To economically save objects of random, often large size on the discrete space on the tapes, it is necessary to optimize it. The archive temporarily gathers on its own disc buffer, objects supplied by The Repository that are ready for archiving. Selection algorithm choosing from this pool always the biggest objects that fit to tape capacity allow to use as much of the tape capacity as possible. It must be noted that each object in the Repository is smaller than tape capacity.

Figure 2. Tape capacity optimization

However this algorithm may cause some objects to wait in the buffer for a very long time. To avoid this selection algorithm in first pass selects form objects that stay too long in the buffer and in the second pass selects from the others.

Contradictory parameters such as the maximum allowable time of the object in the buffer and the minimal tape capacity loss will be fine-tuned on the basis of statistics. After achieving at optimal thresholds, the tape is written only once and never changed. This approach ensures maximization of write speed and durability of the tape. This also allows to overcome the incompatibility between WORM and LTFS by switching write protection tab on the tape cartridge further improving safety of the archive.

The Archive System keeps many independent tape archive replicas in different locations, capable to connect in simple way to more locations, controls their consistency, verifies checksums and carries repairs in case of damage. of automatic May serve different configured clients.

The result is versatile solution currently working in production and already stored more than half of petabyte of unique data. Each object stored in The archive on tape may be read in any LTO7 drive with the use of open source driver and contains human readable contents.

4. REFERENCES

[1] Buffington, J. 2016. *Analyzing the Economic Value of LTO Tape for Long-term Data Retention.* http://www.lto.org/wp-content/uploads/2014/06/ESG-WP-LTO-EVV-Feb_2016.pdf

[2] *ISO/IEC 20919:2016 Information technology – Linear Tape File System (LTFS) Format Specification* (First edition) https://www.iso.org/obp/ui/#iso:std:iso-iec:20919:ed-1:v1:en

[3] http://www.loc.gov/standards/mets/mets-schemadocs.html

[4] http://www.oclc.org/research/publications/library/2000/lavoie-oais.html

Digitisation And Long Term Preservation In a Distributed Environment – The Approach Of The Hungarian Comprehensive Library Platform

Miklós Lendvay, IT director, National Széchényi Library, Budapest, Hungary

Abstract

Renewing it's IT infrastructure and the complete software environment the National Széchényi Library took a leading role in defining a nationwide Comprehensive and Distributed Library System, which is not only taking care of the traditional library functions, but also gives room for a number of areas of inter-institutional cooperation in the fields of digitisation, digital preservation and webarchiving. Also the increasing proportion of digitally born contents in the collections makes this new collaborative approach necessary, involving also scientists outside of institutions, using the means of crowdsourcing etc.

Only distributed systems can ensure, that the libraries can cope with the increasing workload of cataloging, classification and linking of content in a semantic (meaningful) context. Therefore the specification and the definition of this nationwide system had to be the result of interlibrary cooperation, and so will be the implementation, running and support of this platform a joint effort. The linked namespaces give the cooperation even a bigger scale: cultural and non-cultural institutes will work together, sharing and linking their data. Sharing the infrastructure, sharing the software, sharing the data, sharing the effort. A fine balance has to be found in Long Term Preservation: the number of copies, the formats, the readability of the formats etc. are challenges, which should be addressed together in the re-forming library community.

Full version of the article

In historical times and in modern days alike it was and is a primary task of national libraries, to collect the wisdom of a community, preserve it for an infinite time, digest it, and promote it to a wider audience. The scope and the form of this library process varied drastically during the few centuries-decades since libraries exists, but it took definitely the wildest and probably the most unexpected turn in the XXIst century.

The knowledge and the information has been always a big value, and provides it's owner a huge power, and the potential of influence on many areas of life. Only the chosen, the privileged people could once access the strictly protected materials, a very few could understand the signs, could read and write, the language of knowledge was for many countries a foreign language, i.e. for commonfolk a coded language – the Latin language prevailed for centuries. There were many obstacles to overcome, if someone wanted to access the information – which was in earlier times initiation knowledge, secret religious content. The new Testament brought a different approach to the privilege-tradition: what was known, had to be publicly revealed – the forgotten knowledge had to be re-understood, re-digested again, in a new form, in a new way – for everyone, by means of speech and means of written materials.

And the modern times took a further step, made all the secret knowledge public, first in books and periodicals, then on the world wide web, the esoteric knowledge became exoteric. With this step the other extreme came by: abundance of available information, a vast amount of different sources and statements to choose from, statistical knowledge and big data. What to absorb and whom to believe? – these are burning questions for those seeking profound understanding of the world. The human being became free, free of tradition and foretold values, the path is not paved any more, the freedom of choice is given for the wide wealthy enough public. This is the base situation for a national library today in Europe: to serve the free human being in his quest, providing all the necessary information without selection and moral judgement. "All" meant earlier a scope of materials, what one could oversee and a library could cope with; nowadays "all" means much more content, than what would be digestable for a library or even a library community.

We see a huge change in all aspects of a national library's common task: in the proud motto of a modern National Library: "HERE WE HAVE EVERYTHING" the meaning of each and every world had a fundamental change. EVERYTHING went from the paper-based manuscripts and printed materials to a vast amount of electronic documents, in the widest variety of technical formats: documents, digitized objects, sound and video recordings, the row is endless. The formats are ever evolving, dying – and

not easily understood any more – and new formats are continually being born. “HERE” went from the physical building and stock to an undefined cloud of infinite dimensions, the reading rooms in the building went to a widely open, worldwide accessible knowledge-well, where information is available within a fraction of milliseconds. “WE” went from a securely guarded institution to a community of libraries, other institutions and private persons, contributing to the NL collections remotely. “HAVE” meant earlier guarding, maintaining and taking care of the physical objects, keeping the humidity constant and the light level low; this has been extensively extended by the challenge of keeping the contents of the electronic resources, transferring them into actual formats, which can be decoded by the newest technology, with the newest hardware as well. By doing all of this a library always tries to maintain the content, but if it is possible, the form as well; taking the effort of keeping the vinyl LP-s playable with vinyl LP player, cassettes, CD-s, floppys with the appropriate devices. This means to open a sort of device museum as well, so the devices become an integral part of the collection.

New challenges: vulnerability of information, protection of copyrighted materials and personal data (with many new aspects, considering the GDPR being introduced in the EU in spring 2018), conversion of content from an outdated format into a newly developed one, viruses, hackers – all the challenges of the digital era, of information technology.

With the growing number of materials it is not possible any more to collect everything in one library. The increase in personnel has not only it’s financial, space etc. limitations, but would not make use of a co-operational possibilities of modern infrastructure and sharing of labor.

So here we are: a vast amount if information from reliable and less reliable sources, printed and digitized, digitally born or web-born cultural materials.

The biggest impact on the life of a library has the revolution of access in information technology. ACCESS. This does not mean only the capacity to access materials, contents, but also to use central software, shared workflows, with countless possibilities of communication and information exchange. Every participant – embedded in an institution, a freelancer or an individual expert – with an internet access has the possibility with sufficient authorities to contribute to the process of collection, cataloguing, classification, enriching, providing copyright information etc. – on all levels of the library work.

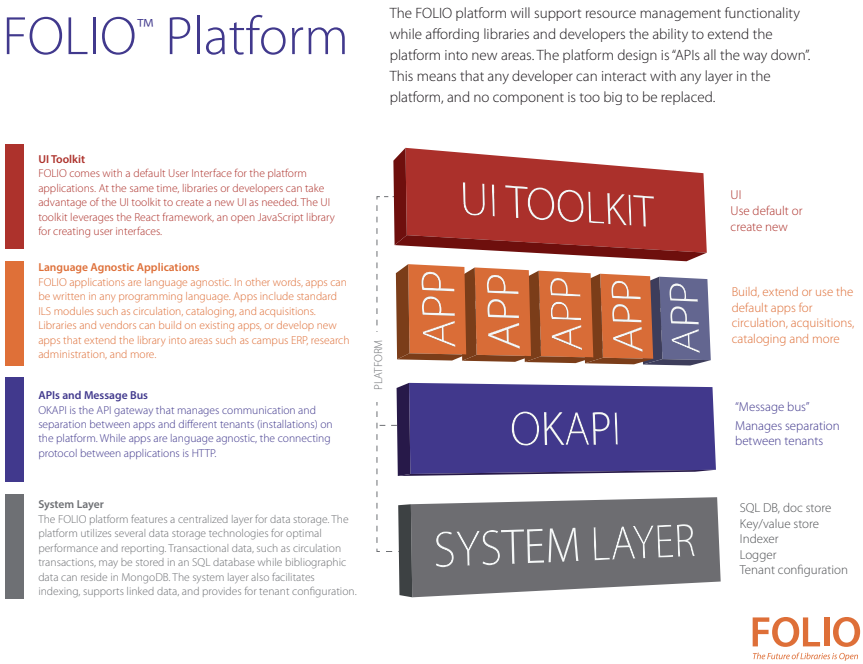


Figure 1: FOLIO

The development of information technology brought not only the challenges of mountains of materials to be classified, catalogued, processed, digitized, but gives also the necessary tools to share the work between libraries, institutions, private researchers, crowdsourcing etc. In a spiritual sense social democratic world, where everyone can consume, and everyone can contribute.

There is no such an integrated library platform in the world, which could fulfill this requirement. To achieve such a goal, there is a need to create a completely modern approach, and put a system together, which takes into consideration all the above ideas and rules.

So on the one hand the National Széchényi Library has joined the initiative FOLIO, which means: The Future Of Libraries Is Open. This initiative is an open community, which set himself the task to develop a truly modular, programming language agnostic cooperation framework, where every participating institution can contribute with self

designed or / and developed modules, can make use of the OKAPI communication layer, and make his own creation accessible and reusable for anyone. A free, open source software, which is modern in the design, free of charge, and is based on the most important achievement of our age: the free collaboration of free participants, providing free choice for everyone. The most needed improvement of such an approach was that the librarian community can set up the rules, how a library platform should look like, and the developer community can create the necessary technical tools, modules and communication channels, to make the community approach work. And again the librarians and all non-librarian contributors can fill this framework with content, with data of all kinds.

On the other hand the biggest and most influential Hungarian libraries gathered at the end of 2016, and expressed their wish to have a common, new generation, cloud based library platform. These libraries have issued a letter of intent about their intention. In the same year the government of Hungary has provided funds for the National Library, to renew the complete information technology, hardware and software infrastructure alike.

These two streams united: the work with the FOLIO community and the work with the Hungarian library community resulted in an intensive cooperation in creating a platform specification, defining, what the Hungarian community wants. It was obvious, that the National Library should not only renew it's own hardware and it's own software, but should make a cloud based server and storage infrastructure and provide a software solution hosted in this cloud, providing a base for all kinds of libraries: public, university, religious, etc., beside the unique and special functions of the national library: providing the national bibliography, being the ISBN and ISSN agency of the country, collecting all the legal deposits in the printed and in the digital realm, web-harvesting, creating the national namespace etc.

The community is open to every library in Hungary, and quite a number of libraries have entered through the open door. There is a substantial effort put into this system from a number of libraries. They worked intensely on the design and definition, and came up with a high level specification of a library platform and tender specification, with a 1.500 item long detailed system requirement list. The National Library has published the European tender, and hopefully after a successful process, in the year 2018 we can start creating this comprehensive library system.

In this process of renewing it's complete IT processes, the Hungarian National Library is and giving an open and shared answer for the challenges of digital preservation and

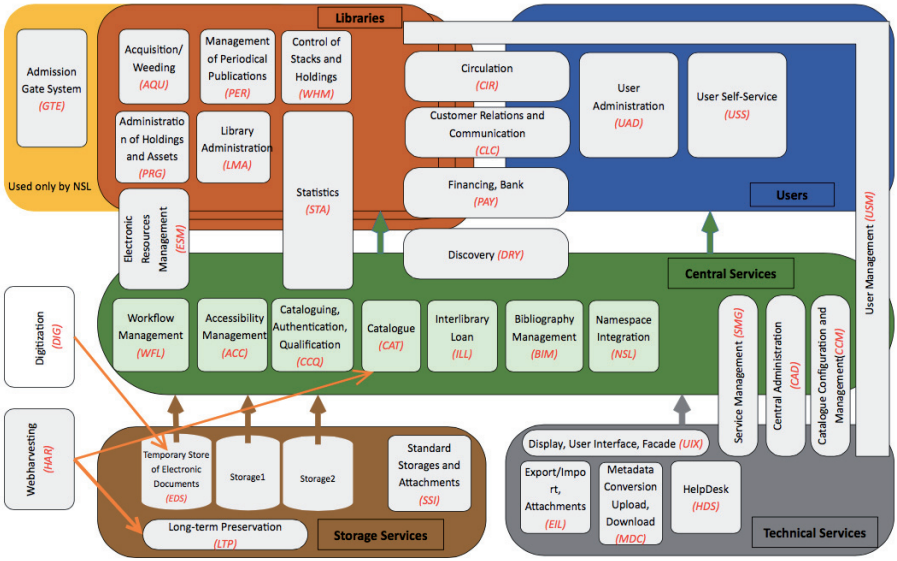


Figure 2: NATIONAL LIBRARY PLATFORM

promotion. At the end of the renewal we wish to achieve a completely new and modern infrastructure, a cloud based server-cluster, central storage, with a wide broadband access to the network of libraries and to partner's infrastructures. The very fundament of this is a community approach of shared definition of the model, shared effort in cataloguing and processing of data, shared data, shared software and shared workflows.

There will be a central software accessible for all participants. The depth and level of collaboration can be individually chosen. For libraries it is free to use the software as their main library platform, maintaining all their data in it, sharing the common data, keeping the sensitive data private and protected. This platform software is used by many libraries in parallel, a real multitenant system. A more loose way of cooperation is, as libraries use their stand-alone and independent traditional Integrated Library Systems, and use the central data, and contribute to the central information. This can include the synchronization of the database, or a regular upload and / or download process. An even less formal contribution could be the data content kept outside the system, pointing to it with a link from the central database.

The identification of things has a significance in linking the data. The new type of data relations make possible a meaningful connection between remote areas. Entities are

labeled with unique identifiers, and these makes the relations concrete and unique. The big disadvantage of MARC is, that this description is a linear one. For each item the around two thousand fields and subfields of MARC are used to describe a library unit, a book, newspaper etc. It is using strings instead of identifiers, and through the possible duplications of strings (in case of same string = different entity) this prevents us from having reliable and trustworthy connections. The library world have realized this, and have introduced the RDA (Resource Description and Access) cataloging standard and the exchange format BIBFRAME (Bibliographic Framework) has been developed to exchange and store data in a format, where well defined links between data are possible.

The big advantage of this is that the entities have to be described only once, and all the users (libraries and end-users) can access, use and reuse it. The multiple effort of cataloguing everything many times in many libraries can be eliminated. The nonlinear BIBFRAME model is saving a lof of effort, and the cooperation and sharing between libraries contributes even more to this.

Earlier the publications have been rare, handwritten books, codices etc. With the development of printing techniques many copies have been produced. With the publication of digital (digitized or digitally born) versions of works one copy is enough, if the access is provided – many can access the same copy. This is similar development like in cataloguing: with MARC, one describes the same work many times, the same author, the same geographic place etc., and with the creation of persistent and unique identifiers in central places and with the access to those descriptions in the centralized systems the duplication is obsolete. The work itself is decentralized, but the identifiers are in a very central place.

A very important part of our project is the introduction of the RDA cataloging, the BIBFRAME data format and the FRBR (Functional Requirements for Bibliographic Records) conceptual entity – relationship model (developed by the International Federation of Library Association). The linked data model approach will be supported by an even wider cooperation between institutions: the National Namespace, which will consist of uniquely identified entities: persons, authors, geographic places, institutions, works etc. In the building of the namespaces (ontologies) not only libraries take part, but also governmental and geographical institutions, cartography map providers, museums, archives and many more are involved.

The new library platform has a central part for identification: identification of users, of digital objects, entities like authors, geographical places, corporations etc. The de-

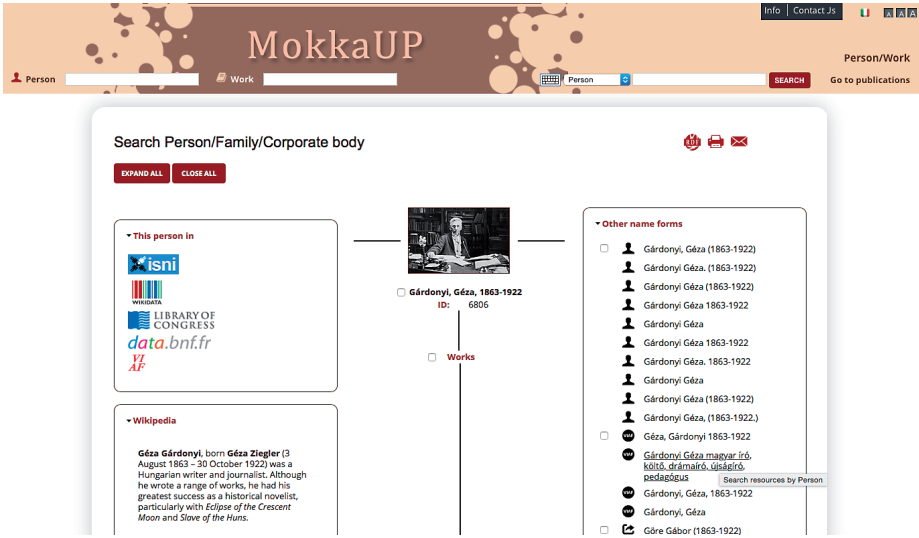


Figure 3: MOKKA UP

scription of an entity in a central place with a unique identifier has countless advantages, and is the base for creating reliable semantic (meaningful) connections between data – this is very important in providing access for the wider public to the collections. In the everyday library work this saves a lot of effort in cataloguing, as authors, titles, works, manifestations are described only once, and the community can freely use them, and enrich it with descriptive data and additional, related information. No copy cataloging needed any more!

To see how the FRBR / BIBFRAME structure works in practice, we have made a project with the Italian company @cult, and converted the Hungarian common catalogue into BIBFRAME format. This made obvious, that there are many duplications, errors in the database. The data have been collected since 2003 from many libraries following many different cataloging habits. The unified new structure gives the libraries a chance, to correct their erroneous data, and merge unnecessary duplicates.

In the cooperation model the participating libraries will provide not only the catalogue of their printed collections, but also their digital assets into the new system. By law the legal deposit for electronic resources – not only the e-books, but also e-maps, e-journals– have to be submitted to the national library. Also by law, whatever gets digitized in the libraries, the digital file has to be handed in.

The collection, the cataloguing and the processing will be realized in a shared environment. This means, that there will be less duplications, as the already existing information can be used and reused, without duplicating it! Everyone can add to the existing description, and can create a new entity, if he deems it necessary. Multiple entries on purpose are also allowed, not only for catalogues, but different digitized versions of the same edition are possible – this is treated in the system as a separate item. In each case the information is provided, who is giving the "statement" about the entity. No automatic, machine made data manipulation will be done in the background, any duplication and merge of data will be the conscious decision of the librarians – in this process tools will be provided to support the editing.

We plan a digitization environment, where the libraries can centrally specify their plans for digitization, and everyone can see it and avoid duplicated effort, and look for ways of cooperation, sharing the task. The results of digitization will be also visible for everyone immediately in the central catalogue.

Everyone is responsible for his own data, for his own statements. The freedom is given to state anything, but also the responsibility has to be taken by the creator for his statements. To each statement a quality level can be attached, to reflect, on which institutional level is the data verified. A crowd-sourced entry, or a statement done by a contributor from outside of an institution, a low level of quality will be attached. These entries can be proven and qualified by librarians with higher qualification, they can bring the entries to higher quality levels. In the systems also the originality of an entry can be investigated, and the authentication level can be specified accordingly.

The participating libraries are free to handle their own collections in the common software, being responsible for their own data. The skin, the presentation of the data is freely definable in the platform: each library, each collection, each author can use own logos, characteristics, colors etc. on the collection-specific webpages. For the processes within separate projects / institutions individual workflows can be setup, this can be parameterized according the needs of the community working together in the given context.

As a national library, we are obliged to save the data for an infinite time. For the preservation the platform will create well defined units, where the digital objects contain their metadata in them. It takes care for the complete wealth of data, descriptive and technical metadata alike. A separate system will provide the content of the webharvesting, and this will go into the channel of the LTP processes.

In the ideal scenario each individual member of the platform will have the necessary

right to access the own data, and can retrieve earlier copies. For data protection purposes the sensitive and restricted (copyrighted) data have to be encrypted. The central infrastructure will take care of the storing, and through the hierarchy of the storage management the versioning of the preserved data. It has not been decided yet, how much data will go to tapes, and how much will be kept on discs. The tender leaves the technical solution open, and we will know only after the tender, which infrastructure we will build for the appropriate LTP software. We are very open to have a community developed, open source solution, if possible.

Not only the renewal of the complex IT system, but also preservation is a community effort. It is the responsibility of the community, to create identifiable data, digital objects according the strict quality requirements, and to participate in the shared model of preservation. It is really up to us, how we take the responsibility, and preserve our cultural treasures for the future generations!

Everything has it's own value only then, if it reaches those, who are seeking for it. The connection between author's content and the reader has to be established. One of the many bridges can be the library, if it wishes to be. In an epoch, where the readers are moving into a cloud environment of freely available data, we have to make sure that we provide for the users the big advantage of the library: providing classified, identified, and authenticated, reliable data. We are able to create a system, where a user can find trustworthy information in the best possible quality. There is a good advice: "If you want to save it, share it!" Only shared copies are safe! This is the way, how spiritual content can live on, how coded knowledge can become live again in human beings. This is a big challenge, but worth it! We have to make this channel through our library platform transparent, shared, available, free!

I am very glad, that the FOLIO community and the Hungarian community gives this modern community approach a try. I am very proud to be able to take part in it. And as this is an open initiative, I encourage you all, whoever can take part and contribute: learn more about this and join this adventereos work!

Further readings:
FOLIO community: www.folio.org
BIBFRAME model: www.bibframe.org
MOKKA UP project: http://test-mokka-up.oseegenius.it/mokka/clusters
The National Széchényi Library: www.oszk.hu/en
The National Library Platform, webharvesting and digitization project:
http://www.oszk.hu/en/node/3490

Appendix

National Library System Project (NLS/OKR)
Abridged Technical Description for the Procurement of the NLS System
7th of June 2017
Excerpt, the relevant modules, which have an immediate impact and have connection to the Long Term Preservation module

Abbreviations of the modul names used in this section:
ACC = Access Management
ACQ = Acquisition
EDS = Temporary Storage of Electronic Documents
DIG = Digitization
DRY = Search Engine, Discovery
HAR = Web Harvesting
LTP = Long Term Preservation
SSI = Standard Storages and Their Conduits
WFL = Workflow Organisation

5.3.2 Digitization (DIG)

In the National Széchényi Library a remarkable volume of digitization activity has been carried out over the past years, but simultaneously with the establishment and buildup of the National Library Platform (NLP) there are developments in the area of digitization as well which will result in a significant increase of digitizing capacity.

Digitization as a self-standing module cannot be identified, yet on account of its insertion and adaptation into the process as well as its integration into partial elements it is important to highlight the seminal correspondences.

Within the National Library Platform (NLP) it is first and foremost the acquisition and CCQ modules where the documents described in the catalog can be earmarked and flagged for digitization demand. This demand can be individual (as a service required or perhaps paid for by the end user) or group-based (eg. on bibliographic grounds or according to the hit list of DRY search). It is expected of the National Library Platform (NLP) that the state of documents thus earmarked for digitization can be traced during the process management of outside (non-NLP) application. At the end of the digitization process there are functions for linking the produced digital objects to the catalog entries.

The libraries using the National Library Platform (NLP) are allowed to see each other's plans, their digitizations in progress and their digitized copies. To a catalog entry several digital contents can be assigned, insofar as more than one digitized version is produced either in the same institution or in other institutions. These versions can be regarded as a kind of digital copy.

On the basis of the above the digitization is to be carried out by inception in a National Library Platform (NLP) process and, following a technical cycle of measures, it must return to the NLP process so that the necessary settings for linking can be fixed in the catalog. This process can be executed even between libraries as well.

Within the National Library Platform the digitization plans of libraries are to be recorded in a simple environment of project definition and this can be queried for all the libraries of the NLP. Thus, even collaborative digitization projects can be designed in the future

5.4.2 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 mangement 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 are 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.

5.7.1 Temporary Storage of Electronic Documents (EDS)

In this storage are placed those files which arrive either from digitization or via other channels, but with the arrival the related workflows have not been terminated yet. The fast-access temporary stores linked to the workflow administration make the electronic document available during execution of the individual work phases (eg, arrival check-in, processing) pinned to authorization, and with remote access as well and a minimal need of mandatory metadata. In the process it is possible to assign qualification and authentication levels to documents, automatically during the procedures of upload, load and machine processing or by staff members too with tools and surfaces adequately shaped for this purpose. Upon execution of the proper administration, processing, cataloging tasks the objects will get into the archive storage or the deposit of service copies warranting long-term preservation.

The temporary provisional storage does not separate physically from the database and the sorting layers it is simply a logical view of the complex procedures of the qualification and authentication of data.

5.7.2 Long-Term Preservation (LTP)

The software effectuating the function of long-term preservation is part of the range of this Request for Participation. As a result of the realization of the function it is expected that the NSZL can satisfy its obligations of long-term preservation specified in legal rules (Act 140/1997 On the antique institutions, the public library supply and public learning 134. § 5; 30/2014 (IV.10.) EMMI Decree 8 § (1) 4). The hardware infrastructure, necessary for long-term preservation, is supplied by the tender announcer.

The task of the system warranting long-term preservation is to maintain the digital documents and data of the collection of the National Széchényi Library in the long run.

In the NLP, long-term preservation is the duty of the National Széchényi Library, so this function is not to be attained for other libraries by the Vendor.

The types of digital objects to be treated in the LTP module: still, motion picture, text (of binary coding [eg, doc, ppt, xls stb.], (html, xml, ALTO xml, TEI etc.) voice, audio, 3D and the combination of these. The most frequently used formats: JPEG2000, JPG, HTML, XML, PDF, PDF/A, e-pub, VAW, MP3 etc. The preservation of backups and saves generated during web harvesting is the duty of this module as well.

The system of the NLP uses standard metadata and de facto digital formats. The treatment of digital objects in the NLP system takes place according the OAIS model (ISO 14721:2003).

The creation of a software-based solution supporting the multiplication and storage necessary for the long-term preservation of digital objects is also an expectation from the module. It must enable the identification, integrity, software-supported readability of the digital documents and objects the integrity of its structure and the execution of operations related to the objects (control, supervision, migration, conversion, compression, etc.).

The handling of the LTP module is performed by staff members and/or programs with the appropriate privilege. It is guaranteed that the state prior to the execution of an operation can be restored and recovered, even concerning multiple operations as well.

The handling of the metadata necessary for long-term preservation is to be solved by the vendor, eg. on the basis of Preservation Metadata: Implementation Strategies [PREMIS], Metadata Encoding and Transmission Standard [METS] or other de facto standards.

The module can handle the metadata pertaining to the history of digital objects (eg, migration, conversion, compression, control etc.).

5.7.3 Standard Storages and their Conduits (SSI)

The component must enable the coupling to repositories which are not in the pooled storages of the NLP and to which direct access is to be secured via the catalog entry. Therefore, the NLP is to communicate with the software programs of repositories (Eprints, Dspace, Jadox) to ensure the transfer of metadata between repository and catalog, to yield the documents' access data to the query system. It must collaborate

with other systems (eg, Open Journal Systems, Magyar Tudományos Művek Tára – MTMT).

5.8.2 Web Harvesting (HAR)

The functional content of web harvesting is not the subject of this acquisition, while the access and treatment in the NLP of the web archive arising from the result thereof must be realized the same way as the treatment of images, videos, etc.

The self-contained project of web harvest development, which is going to be effected separately, aims to create the concept, structural frames and information infrastructure of a would-be Hungarian internet archive (webarchive) and to set up a test archive.

The ultimate goal of the web harvest development project pointing far beyond the NLP project is the creation of a system which, in addition to the task of preserving Hungarian and Hungary-related cultural heritage appearing on the internet in the long run, will serve the requirements of education, scientific research, state organizations, business sphere and some internet users.

The public service of metadata and fulltext search engine of the web archive and the service of archived web pages within the NSZL and the NSZL and NAVA points on dedicated computers must be coupled up to the NLP as any other storage of digital objects. In the course of the handling of the NLP catalog an option must be present for supplying the html-based digital objects placed in the web archive with metadata (for cataloging), or for transferring metadata into the catalog by means of the MDC module and with the help of these metadata the queries of the module must expand to include the web archive. The running of fulltext retrieval by special software and the delivery of the results while retaining the storage of data in the web archive is a sufficient and acceptable solution.

The web archive is substantially related to the access managing module (ACC) ensuring access to the contents by privileges registered in the system. The maintenance of a part of the privileges takes place in the web archive, and this is related to the ACC module. It must be able to regulate belated or delayed accesses related to the age of the archived contents.

The secure storage of archived oversize files and their readability is solved by its integrated relation to the long-term preservation system (LTP).

Projekt ArcLib – příprava metodik a vývoj open source řešení pro dlouhodobou archivaci digitálních dokumentů

Ing. Martin Lhoták, Knihovna AV ČR, v. v. i.

Abstrakt

Článek představuje projekt, jehož cílem je vytvoření komplexního LTP (Long Term Preservation) řešení ARCLib na bázi open source softwaru, které využije volně dostupné nástroje a systémy. Součástí projektu a jeho dalším významným výstupem je vytvoření metodiky na dlouhodobou logickou ochranu digitálních dat zohledňující mezinárodní standardy v této oblasti a informační systémy využívané v českých knihovnách. Tato metodika, která již byla předána k certifikaci, bude v článku podrobněji představena. Současně bude připravena metodika a pro fyzické ukládání dat a zajištění bit-level ochrany. Nové LTP řešení ARCLib umožní zajištění dlouhodobé ochrany digitálních dat s podporou OAIS modelu v knihovnách různé velikosti a bude volně dostupnou variantou ke komerčním softwarovým řešením, jejichž nasazení je většinou doménou velkých institucí typu národních knihoven a národních archivů. Kromě knihoven se může stát vhodnou volbou pro další paměťové instituce – muzea, galerie a archivy.

Úvod

Projekt ArcLib reaguje na potřebu paměťových institucí a zejména knihoven zajistit dlouhodobé uchování digitálních dokumentů. Součástí projektu je příprava metodických materiálů i technického řešení, přičemž vše bude volně dostupné – metodiky formou open access a vyvinuté softwarové nástroje jako open source.

Na řešení projektu vedeném Knihovnou AV ČR, v. v. i, spolupracují Masarykova univerzita, Národní knihovna ČR a Moravská zemská knihovna v Brně. Projekt je pod označením DG16P02R044 řešen v období 2016-2020 s finanční podporou od Ministerstva kultury ČR v rámci dotačního programu NAKI II.

Cíle řešení projektu

Cílem projektu je vytvoření komplexního LTP (Long Term Preservation) řešení ARCLib na bázi open source, které využije volně dostupné nástroje a systémy. Součástí projektu a jeho dalším významným výstupem je vytvoření metodiky na dlouhodobou logickou ochranu digitálních dat zohledňující mezinárodní standardy v této oblasti (referenční model OAIS – ČSN ISO 14721 a ČSN ISO 16363) a systémy využívané pro vytváření a zpřístupňování digitálních dat v českých knihovnách. Současně bude připravena metodika a řešení pro fyzické ukládání dat a zajištění bit-level ochrany. Funkčnost celého řešení bude ověřena v praxi formou poloprovozu minimálně v jedné ze zapojených institucí. Pro potřeby projektu bude mimo jiné využit open source systém Archivematica. Nové LTP řešení ARCLib umožní zajištění dlouhodobé ochrany digitálních dat s kompletní podporou OAIS modelu v knihovnách různé velikosti a bude volně dostupnou variantou ke komerčním řešením, jejichž nasazení je většinou doménou velkých institucí typu národních knihoven a národních archivů. Bude zabezpečena interoperabilita s LTP systémem Národní knihovny ČR, která umožní obousměrnou výměnu archivačních balíčků pro dlouhodobou ochranu. Spolupráce obou archivačních řešení bude významně zvyšovat stupeň ochrany uložených digitálních dat. Nové řešení ARCLib může být v budoucnu zálohou a případnou alternativou i pro LTP úložiště NK ČR, která existencí záložního řešení naplní požadavky ve smyslu normy ČSN ISO 16363. ARCLib bude otevřeným řešením, které v případě potřeby umožní připojení dalších systémů do archivačního procesu. Kromě knihoven se tak může stát vhodnou volbou např. i pro další paměťové instituce – muzea, galerie a archivy.

Popis jednotlivých cílů

1) Vývoj komplexního LTP (Long Tem Preservation) open source řešení ARCLib

V knihovnách České republiky trvale narůstá množství projektů a aplikací generujících velké objemy digitálních dat. Probíhají rozsáhlé digitalizační projekty a stále více dat vzniká přímo v digitální podobě (born-digital). Pro velkou část těchto dat je ne-

zbytné zajistit jejich dlouhodobou ochranu a přístupnost (Long-term Digital Preservation, LTP). Je třeba zajistit jak tzv. bit-level ochranu dat (zabezpečení před fyzickou ztrátou, změnou či havárií digitálních souborů a nosičů) tak logickou ochranu (ochrana před nepříznivými dopady změn a zastarávání informačních technologií a datových formátů na dostupnost a použitelnost digitální informace).

Problematika dlouhodobé ochrany digitálních dat (LTP – Long-Term Preservation nebo DP – Digital Preservation) byla až donedávna výhradní doménou velkých institucí typu národních knihoven či národních archivů, které disponovaly potřebnými mandáty, financemi a expertními zdroji. Tyto instituce se typicky zaměřily na budování komplexních na míru vyvinutých řešení postavených zejména na komerčních systémech. Pokroky v oblasti teorie a praxe digitální ochrany spolu s rostoucími potřebami řešit dlouhodobou archivaci digitálních dat i v menších institucích vedly k poznání, že i s omezenými zdroji lze začít vytvářet vlastní řešení s využitím volně dostupného softwaru (viz např. projekt POWRR – Preserving Digital Objects With Restricted Resources, http://commons.lib.niu.edu/handle/10843/13610).

Cílem projektu je vytvoření volně dostupného archivačního systému, který bude respektovat národní i mezinárodní standardy. Pro potřeby projektu bude využit open source systém Archivematica, který je dynamicky rozvíjen a nasazován v řadě projektů po celém světě. Archivematica však neřeší všechny funkční entity dle modelu OAIS, ale zaměřuje se jen na kritické archivační funkce (transfer, příjem, tvorba informačních balíčků SIP/AIP/DIP).

Nové archivační řešení ARCLib bude vyhovovat požadavkům odvozeným z funkčního a informačního modelu standardu OAIS, tj. mělo by ochraňovat informační obsah v balících AIP se všemi OAIS metadaty a mělo by disponovat nástroji pro podporu všech funkčních celků OAIS (OAIS functional entities) včetně celku "plánování uchovávání" (preservation planning). Komunita uživatelů systému k tomu pak bude společně udržovat znalostní základnu potřebnou ke kvalifikovaným rozhodnutím při dlouhodobém uchovávání informačního obsahu ve vyvinutém systému – databázi formátů, pravidel a služeb, migračních cest, nástrojů – a vykonávat funkce požadované standardem OAIS v oblasti plánování uchovávání.

ARCLib bude kompatibilní s komerčním řešením Národní knihovny ČR a umožní předávání archivních balíčků AIP mezi instancemi nově vyvinutého systému navzájem a se systémem LTP v NK, a naopak. Z hlediska modelu OAIS se jedná o možnost vytvoření sítě spolupracujících OAIS archivů propojených standardem pro "repository exchange package" (např. jako http://wiki.fcla.edu/TIPR); výstupní balíček DIP z jed-

noho systému by měl sloužit jako vstupní balíček SIP dalších systémů (a systému NK), a naopak. Interoperabilita umožňující oboustrannou výměnu archivačních dat s komerčním LTP řešením v Národní knihovně významně zvýší stupeň zabezpečení archivovaných dat v České republice. Zároveň může NK ČR naplnit požadavek archivačních standardů na existenci exit strategie.

Vstupem pro archivační řešení ARCLib budou data ze všech majoritně využívaných softwarových řešení pro výrobu, zpřístupnění a ukládání knihovních digitálních dokumentů v České republice. Zejména se jedná o digitální dokumenty ze systémů:

- Kramerius – systém pro zpřístupnění digitálních dokumentů; využívaný ve většině velkých knihoven v ČR
- ProArc – systém pro výrobu digitálních dokumentů; využívá např. Knihovna AV ČR, SVK Hradec Králové
- DSpace – repozitář využívaný zejména na univerzitách jako přístupový systém jak pro digitalizované sbírky tak pro nově vznikající digitální dokumenty (archivy vysokoškolských prací, institucionální repozitáře vědeckých publikací a výzkumných dat); využívá např. Masarykova univerzita, VŠB-TU v Ostravě, Univerzita Pardubice, Univerzita Tomáše Bati ve Zlíně, ČVUT Praha a další

K dalším systémům, pro něž by bylo vhodné zajistit napojení na archivační řešení, patří repozitář Invenio, který je používán v Národní technické knihovně pro Národní úložiště šedé literatury.

Systémy sloužící pro zpřístupnění (Kramerius, DSpace, Invenio) budou zároveň cílovými systémy pro balíčky DIP sloužící k dalšímu šíření dlouhodobě uchovávaných informací koncovým uživatelům.

ARCLib bude otevřeným řešením, které v případě potřeby umožní připojení dalších systémů do archivačního procesu. Kromě knihoven se tak může stát vhodnou volbou např. i pro další paměťové instituce – muzea, galerie a archivy.

2) Vytvoření metodiky pro dlouhodobou logickou ochranu digitálních dat pro české prostředí s ohledem na mezinárodní standardy (zejména referenční model OAIS – ČSN ISO 14721 a ČSN ISO 16363)

Metodika, kterou bude certifikovat Ministerstvo kultury ČR, bude jasně definovat způsob zajištění dlouhodobé logické ochrany digitálních dat zohledňující mezinárodní standardy v této oblasti (referenční model OAIS – ČSN ISO 14721 a ČSN ISO 16363)

a systémy využívané pro vytváření a zpřístupňování digitálních dat v českých knihovnách.

České knihovny i další paměťové instituce tak získají srozumitelný postup pro zavedení a provozování dlouhodobého úložiště digitálních dokumentů a dat různých formátů.

Podrobnější popis nově vytvořené metodiky je v samostatné kapitole tohoto článku.

3) Vytvoření metodiky a návrh řešení pro fyzické ukládání velkého množství dat a zajištění bit-level ochrany pro potřeby dlouhodobé archivace

Výstupem bude metodika certifikovaná Ministerstvem kultury ČR pro fyzické ukládání dat a bit-level ochranu v rámci systému ARCLib pro potřeby dlouhodobé archivace digitálních dat a dokumentů. Součástí této metodiky bude popis základních nároků na úložiště, které lze za účelem dlouhodobého ukládání dat spolu s bit-level ochranou využít.

Navržená metodika musí zvážit a eliminovat rizika, která ohrožují datová úložiště (selhání hardware, neúmyslnou chybu obsluhy, úmyslný útok obsluhy nebo jiného subjektu, přírodní katastrofy, ozbrojené konflikty, legislativní omezení nakládání s daty ukládanými na určitém území apod.) a stanovit vhodné postupy pro minimalizaci škod způsobných působením takových událostí – ukládání identických kopií dat ve více geograficky oddělených lokalitách na různých typech úložišť spravovaných různými skupinami osob, ovšem při zajištění pravidelných kontrol dostupnosti a integrity dat. Nedílnou součástí politiky zacházení s daty musí být její pravidelné revize a úpravy dle změněných okolností v průběhu času. Popsané technické řešení musí být v souladu s politikou definovanými požadavky na exit strategie (export všech dat ve vhodném tvaru pro přenos do jiných/novějších systémů).

Navržené řešení musí být dobře škálovatelné (menší i velmi velké objemy dat, rozvoj systému s ohledem na počet zapojených účastníků), s dobrou propustností (technicky řešitelné např. hierarchickým uložením dat s rychlým online-přístupem často využívaného menšího objemu dat versus off-line uložení velkého množství dat, které však znamená velké latence při vybavování dat) a musí umožňovat použití i více nezávislých řešení podle specifických potřeb jednotlivých institucí. Řešení se dále musí vypořádat s některými základními omezeními, která jsou specifická pro řadu typů úložišť (např. jak datová úložiště CESNETu, tak i úložiště založená na cloudech) – kupříkladu obtíže a omezení při ukládání příliš velkého množství malých souborů aj.

4) Ověření v praxi formou poloprovozu

Archivační řešení ARCLib bude formou poloprovozu nasazeno a ověřeno v Knihovně AV ČR. Budou na něm uloženy např. digitální dokumenty pořízené ze sbírek a fondů ústavů Akademie věd ČR a knihovny Národního muzea.

Rozbor stavu řešení problému v ČR a v zahraničí

V České republice již od druhé poloviny 90. let probíhá digitalizace a zpřístupňování knihovních dokumentů v digitální podobě. Mezi nejaktivnější patřily od počátku Národní knihovna, Moravská zemská knihovna, Knihovna Akademie věd a některé vysokoškolské knihovny (např. na Masarykově univerzitě)[1]. V posledních letech se tato aktivita významně zvýšila také v krajských knihovnách díky dotacím z evropských fondů. Dlouhodobá ochrana digitálních dokumentů je přímo řešena pouze v projektu Národní digitální knihovna (NDK) [2], ve kterém používá Národní knihovna ČR společně s Moravskou zemskou knihovnou v Brně pro dlouhodobou archivaci komerční systém. V ostatních knihovnách není v současné době žádné archivační řešení implementováno. V Knihovně AV ČR je vyvíjeno řešení pro produkci digitálních dokumentů ProArc, které bude současně zajišťovat určité archivační funkce (vytvoření SIP balíčku pro vložení do archivačního systému) nebude však k dispozici řešení plně pokrývající všechny části standardu pro archivaci dle referenčního modelu OAIS. Knihovny v České republice mají k dispozici standardy pro archivaci ČSN ISO 14721 a ČSN ISO 16363, chybí však jasná prováděcí metodika s ohledem na české prostředí a v něm používané systémy a formáty. Současně nejsou k dispozici snadno dostupná ucelená řešení pro archivaci, jakým je v případě zpřístupnění například open source systém Kramerius. Za velkou výhodu z hlediska výchozích podmínek dlouhodobé archivace lze považovat existenci standardního formátu NDK, který je při vytváření digitálních dokumentů následován ve všech významných digitalizačních projektech i v projektech menšího rozsahu. Standardní balíček NDK je založen na v současnosti obvyklých a doporučovaných formátech metadat (METS, PREMIS, MODS, Dublin Core, MIX a ALTO XML), které vycházejí zejména ze standardů doporučovaných Kongresovou knihovnou a jsou respektovány na mezinárodní úrovni. Základní předpoklady pro dlouhodobou archivaci jsou použitím tohoto standardu zajištěny, stejně tak i možnost sdílení dat mezi jednotlivými digitálními repozitáři vede ke zlepšení podmínek ochrany digitálních dokumentů. Standard NDK je určen pro digitalizované dokumenty a je potřeba pracovat na jeho rozšíření i pro e-born (digital born) dokumenty.

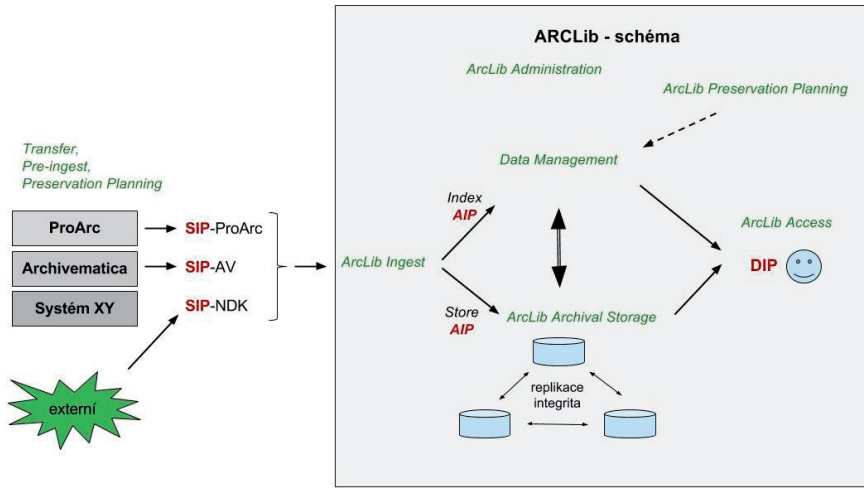
Národní knihovna České republiky v současné době využívá v rámci projektu Národní digitální knihovny pro dlouhodobou archivaci na míru vyvinuté komerční řešení. To slouží zejména pro dokumenty digitalizované Národní knihovnou a neposkytuje obecné řešení pro celou komunitu. Řada dalších dokumentů digitalizovaných v krajských a specializovaných odborných knihovnách se do archivačního řešení Národní knihovny nedostává a odpovědnost za jejich dlouhodobou archivaci zůstává na jednotlivých knihovnách. Všechny knihovny, které ukládají digitální a digitalizovaná data, by měly mít k dispozici volně dostupné softwarové řešení, které jim umožní data ochránit v souladu s požadavky normy OAIS. Archivační řešení v Národní knihovně je v současné době také omezeno několika faktory organizačního i technického rázu. I přesto, že současný vývoj směřuje k odstranění těchto omezení, stále není vyřešen postup (a neexistuje ani návrh jeho řešení) v případě ukončení činnosti úložiště NK ČR tak, jak si žádá norma ČSN ISO 16363 [3]. Z tohoto pohledu řešení ArcLib představuje cestu, jak může NK ČR naplnit požadavek standardů na existenci exit strategie.

V mezinárodním měřítku se setkáváme s různým přístupem k dlouhodobé archivaci digitálních dokumentů, využívána jsou komerční i volně dostupná řešení. V některých zemích probíhá koordinace na národních úrovních, jinde knihovny spolupracují s akademickou sférou při sdílení znalostí a podpoře výzkumu pro řešení dlouhodobé archivace [4].

Z volně dostupných systémů je pro řešení dlouhodobé archivace nejčastěji využíván systém Archivematica, jehož vývoj je veden kanadskou společností Artefactual Systems Inc. Archivematica je vyvíjena v souladu s ISO standardem OAIS (ČSN ISO 14721), neřeší však všechny funkční entity tohoto modelu [5]. Zaměřuje se jen na kritické archivační funkce (transfer, příjem, tvorba informačních balíčků SIP/AIP/DIP). Pro splnění všech parametrů vyžadovaných standardem je potřeba dalšího vývoje. Celé archivační řešení může být složeno z částí vzájemně plnících určité požadavky standardu, přičemž komplexně budou zajišťovat všechny funkce, tak jak je to plánováno v projektu, jehož výstupem bude archivační řešení ARCLib zastřešující a propojující jednotlivé komponenty.

Dosavadní průběh řešení projektu

Klíčovou aktivitou roku 2016 bylo navržení celkové architektury systému ARCLib, která nabídne komplexní řešení, ale přitom současně poskytne i nezbytnou flexibilitu formou zapojení vícero různých v ČR využívaných či připravovaných subsystémů, zejména v oblasti přípravy vstupních balíčků SIP a napojení na fyzická úložiště digi-



Obr. č. 1 – ARCLib schéma

tálních dat. V rámci přípravných prací na zastřešujícím řešením ARCLib byl projektovým týmem připraven popis aplikace a funkční požadavky, které byly podkladem pro věcnou část zadávací dokumentace pro výběrové řízení na dodavatele programátorských prací.

Popis navrhovaného řešení ARCLib vycházející z analýzy projektového týmu:

ARCLib je systém pro logickou a bitovou ochranu digitálních dat navržený v souladu s požadavky odvozenými z ČSN ISO 14721 (OAIS). Dlouhodobým cílem vývoje ARCLib je vytvoření řešení, které institucím umožní implementovat **všechny funkční moduly OAIS** a které bude respektovat požadavky jeho informačního modelu. ARCLib využívá v maximální míře existujících nástrojů, jako jsou ProArc a Archivematica, a to především pro tvorbu SIP balíčků. Připravené SIP balíčky validuje, konvertuje do archivních balíčků (AIP) a ukládá v souladu s OAIS. **Neopakuje to, co už nástroje typu ProArc nebo Archivematica již umí.**

V první fázi vývoje se ARCLib zaměří především na vytvoření modulů **ARCLib Ingest**, **ARCLib Data management** a **ARCLib Archival storage.** Ostatní funkce je nezbytné realizovat i v dalších částech OAIS, především v **ARCLib Administration, ARCLib Access.** Zejména **ARCLib Administration** má řadu funkcí, které souvisí s dalšími funkčními entitami.

ARCLib je dark archive. Není to repozitář určený ke zpřístupňování dokumentů koncovým uživatelům. Nedisponuje prostředky pro zobrazení archivovaných dat (image servery, prohlížeče apod.). Uživateli ARCLib jsou správci archivních digitálních dat, data jsou po exportu používána systémy zpřístupnění případně další systémy digitálních knihoven (DAM systémy). Aktualizace AIP a vytváření nových verzí AIP probíhá z velké části editací dat v externích systémech (ProArc, Dspace) a následným re -Ingestem do ARCLib.

ARCLib AIP ARCLib je systém pro správu archivních balíčků, nejedná se o systém pro koncové uživatele nebo systém pro management popisných metadat. Z toho vychází návrh struktury AIP a návrh funkcí. Vedle SIP (ve struktuře BagIt) ARCLib ukládá a udržuje **jedno XML s metadaty – ARCLib AIP XML.** ARCLib umožní verzování a mazání AIP.

ARCLib Ingest
ARCLib Ingest předpokládá vstup dat v pokročilejší fázi zpracování, tedy data která už mají podobu plnohodnotného balíčku SIP definovaného obsahového standardu vytvořeného systémy, jako jsou ProArc nebo Archivematica, zabalené do kontejneru BagIt.

ARCLib Data management
ARCLib Data management obsahuje informace o AIP balíčcích (resp. jejich částech) uložených v trvalém úložišti v modulu Archival storage. Poskytuje také index a vyhledávací rozhraní. V ideálním případě doplněné o reporting (nad uloženými AIPy i reporting o výkonu zpracování). Z vyhledávacího prostředí data managementu lze vyvolat událost exportu DIP (nyní s tím, že DIP = AIP), jednotlivě nebo hromadně lze prohlížet informace ARCLib AIP XML.

Vyhledávat lze nad popisnými metadaty, administrativními metadaty a technickými metadaty vytvořenými v ARCLib (tedy v rozsahu obsahu ARCLib AIP XML) a zároveň je možné tato metadata editovat.

ARCLib Administration
Modul administrace umožňuje konfigurovat workflow pro zpracování Ingestu a kontroluje infrastrukturu systému ARCLib. Obsahuje registry uživatelů a jejich rolí, provádí se zde nastavení jejich autentizace a další administrační procesy.

ARCLib Archival Storage
Archival Storage je komplexní služba pro zajištění bitové ochrany umožňující využití

replikace dat do více geografických lokalit a využití více technologií ukládání dat. Archival Storage směrem k ARCLib poskytuje Object Storage použitelné přes jednoduché REST rozhraní.

ARCLib Access
Filosofie přístupu pro systém ARCLib předpokládá, že uživatelé potřebují získat zpět vložená data v původním tvaru. Access tedy v první řadě umožní export AIP jako DIP, s tím že obsah AIP a DIP je 1:1. Další případné zpracování pro zpřístupnění koncovým uživatelům nebo aktualizace obsahu AIP (konverze do Krameria, úpravy metadat, změny struktury a obsahu AIP, opakování validace formátů nebo nová extrakce technických metadat) už probíhají v jiných systémech (ProArc, Archivematica). ARCLib je back-end aplikace a není určená pro koncové uživatele. Nevynucuje tedy dodržování nějaké politiky omezující přístup k datům AIP. Metadata k Access Rights jsou součástí dodaného SIPu, a jsou kontrolována při Ingestu, ale nejsou konvertována do ARCLib AIP XML.

ARCLib Preservation planning
Velká část funkcí funkční entity *preservation planning* bude realizována mimo informační systém ARCLib. **Definice a monitorování *designated community*** a **monitorování technologií** jsou činnosti především výzkumné a organizační povahy, a jejich výkon je předmětem zájmu řady komunit. V České republice hraje klíčovou roli v oblasti standardizace v knihovnách Národní knihovna ČR. Část těchto činností vykonávají příslušná oddělení NK ČR, a uživatelé systému ARCLib se mohou řídit jejich doporučeními a vydanými standardy.

V roce 2017 proběhlo výběrové řízení na dodavatele programátorských prací a byla zahájena příprava prvních součástí systému, které jsou průběžně testovány a diskutovány řešitelským týmem. Současně byla v tomto roce připravena Metodika pro dlouhodobou ochranu digitálních dat a předána Ministerstvu kultury ČR k certifikaci.

Metodika pro dlouhodobou logickou ochranu digitálních dat

Cíl metodiky

Metodika pro logickou ochranu digitálních dat předkládá postupy pro dlouhodobé uchovávání těchto dat zejména z knihovních sbírek pomocí softwaru ARCLib. Definu-

je potřebné kroky k naplnění cíle dlouhodobého uchování digitálních dat na úrovni logické ochrany v souladu s postupy doporučenými mezinárodními normami ČSN ISO 14721 a ČSN ISO 16363. Metodika vychází z těchto norem a představuje jejich aplikaci na konkrétním systému dlouhodobého uchovávání. Na tomto místě je třeba zdůraznit, že takový dokument metodické povahy s aplikovanými postupy v českém prostředí dosud chybí.

Logická ochrana digitálních dat je stále ještě poměrně nový pojem, i když její význam radikálně roste. Přináší jiné a koncepčně odlišné postupy, než jaké byly dosud obvyklé při ukládání digitálních dat. Jádrem dlouhodobé ochrany proto zůstávají kvalifikovaní pracovníci využívající mezinárodně osvědčené a zdokumentované postupy ideálně ve spojení s diverzifikovanými technologiemi. Účelem této metodiky je nejen poskytnutí návodu, jak provádět správu digitálních dat v rámci specifického softwarového řešení ARCLib, ale i popis obecných postupů k provádění kroků logické ochrany a definování nároků na personální zajištění chodu repozitáře. Mezi klíčové vlastnosti systémů poskytujících logickou ochranu patří důvěryhodnost, kterou lze zajistit právě dodržování norem a transparentním chodem systému (ve smyslu odborně zajištěné obsluhy a dokumentovaných procesů i softwaru). Toto je ověřováno řadou certifikačních nástrojů. Součástí předkládané metodiky jsou tedy i doporučené postupy, jak věrohodným způsobem provádět certifikaci.

Cílem metodiky je poskytnout návod uživatelům softwarového řešení ARCLib, jak výše zmíněné postupy aplikovat, jak provádět správu dat v systému a hodnotit rizika. Zároveň dokumentuje konkrétní funkce řešení, popisuje způsob uložení dat, jejich identifikaci a strukturu. Právě tato dokumentace je nezbytná pro uznání důvěryhodnosti repozitářů využívajících systém ARCLib. Metodika tak vychází z mezinárodně definovaných požadavků na systémy logické ochrany digitálních dokumentů a převádí je na konkrétní postupy dostupné v rámci vyvinutého řešení. Využívání návodů této metodiky ve všech zmíněných oblastech (provoz řešení, personální a finanční zajištění a příprava na certifikaci) představuje nezbytnou podmínku pro využití softwarového řešení ARCLib. I když jsou postupy logické ochrany obecně popsány, v každém archivačním řešení se jejich aplikace liší na základě odlišného charakteru softwarového systému.

Pravidla a postupy metodiky byly ověřeny jejími autory jak v souvislosti s jinými systémy (např. systém LTP v Národní knihovny ČR, digitální repozitář Univerzity Karlovy a další realizované systémy, se kterými se autoři během své kariéry seznámili), tak s jejich vlastními badatelskými postupy a při vývoji řešení ARCLib.

Vlastní popis metodiky

Metodika je rozdělena do dvou vzájemně provázaných částí a dalších doplňujících úseků. Jádro metodiky je však v úsecích Teoretická a Praktická část, které popisují obecně platné přístupy v oblasti logické ochrany digitálních dat a doporučení, jak tyto postuláty aplikovat v systému ARCLib. Neméně významná je pak podkapitola Implementační část.

V Teoretické části jsou popsány zásady pro budování důvěryhodného dlouhodobého úložiště digitálních dokumentů, je rozebrán koncept funkčních celků OAIS a dále jsou rozebrány další náležitosti budování repozitáře. Nedílnou součástí této podkapitoly je vysvětlení konceptu informačního balíčku a definování možných strategií dlouhodobého uchovávání. Podkapitola je uzavřena doporučením, jak certifikovat repozitář využívající systém ARCLib podle normy DSA. K podmínkám certifikace se váže i dodatek č. 1.

Praktická část se zaměřuje na vlastní systém ARCLib. Text se zaměřuje na popis architektury systému, jeho konceptuální model a na popis jednotlivých částí systému dle funkčních celků OAIS. Druhou podstatnou součástí podkapitoly je specifikace informačního balíčku v systému ARCLIb. Tato specifikace tvoří jeden ze stěžejních výstupů metodiky, jde o originální návrh odpovídající záměrům, s nimiž byl systém navržen. V této části jsou specifikována převzatá i originálně vytvořená metadatová schémata, která tvoří doporučení pro AIP balíček určený k uložení v repozitáři tak, aby nad ním mohly být vykonávány procesy dlouhodobé ochrany. Podkapitola je doplněna doporučeními pro procesní dokumentaci, jež je nezbytná pro vlastní chod repozitáře.

Závěr jádra metodiky je obsažen v Implementační části. Ta přináší provozní doporučení, zejména doporučení pro organizační a personální zajištění provozu systému tak, aby byly splněny podmínky pro důvěryhodné dlouhodobé uložení, doporučení pro finanční plánování a doporučení pro výběr dalších externích nástrojů, které jsou pro plnou funkčnost nezbytné.

Zdůvodnění metodiky

Metodika se zabývá životním cyklem uložených dat, popisuje proces péče o jejich ochranu a toho, jak k dosažení cílů využít software ARCLib. Vychází z doporučení mezinárodních norem ČSN ISO 14721 a ČSN ISO 16363. Jejich ustanovení lokalizuje do českého prostředí a obecné zásady převádí do nově vytvořených konkrétních postu-

pů a doporučení, jak žádaných cílů dosáhnout v instituci, která bude využívat řešení ARCLib. Metodika vychází z výzkumů provedených v projektu NAKI II ARCLib i v dřívějším působení jednotlivých autorů. Spojuje teoretické postuláty a zásady s konkrétní implementací. Novost postupů představených v této metodice spočívá právě v implementaci teoretických postupů v instituci, která využije řešení ARCLib.

Logická ochrana digitálních dat představuje stále ještě poměrně nový pojem a její provádění zůstává spíše v teoretické rovině. Ochrana není automaticky zaručena využitím konkrétního ověřeného nástroje a patřičným hardwarovým vybavením a není ani možné na základě měřitelných parametrů přesně odhadnout dobu, po kterou jsou dokumenty chráněné na základě jednorázových opatření. Nicméně využití konkrétních nástrojů je nezbytnou (i když ne postačující) podmínkou zajištění logické ochrany. Tato metodika tedy nezůstává jen teoretickým dokumentem, který definuje obecné postupy, ale naopak přináší konkrétní návod pro provádění úkonů logické ochrany postavených na softwaru ARCLib. Novost postupů lze hodnotit ve dvou rovinách.

První z těchto rovin je obecná, kdy je metodika pro logickou ochranu digitálních dokumentů určena všem institucím a jejich odborným pracovníkům, kteří mají záměr dlouhodobě uchovávat tyto dokumenty. Popisuje principy péče o dlouhodobé uchovávání, nutné požadavky na dlouhodobá úložiště, pravidla důvěryhodných repozitářů a nároky na provozující instituce. Takto kompletně pojaté doporučení v českém prostředí neexistuje a není samozřejmostí ani na mezinárodním poli. Novost tedy spočívá především v komplexním provázání doporučených postupů.

V druhé rovině lze novost sledovat ve spojení s doporučeními a návody pro uživatele systému ARCLib. Postupy logické ochrany digitálních dokumentů lze provádět na základě obecných norem jen v konkrétním nástroji. Tímto nástrojem je v kontextu předložené metodiky právě ARCLib. Jde o nově vyvinutý nástroj vzniklý na základě poznatků dosažených v rámci řešení projektu NAKI II ARCLib. Úspěšné využití nástroje je možné jen při dodržení zdokumentovaných procesů, které jsou obsaženy v metodice a v kontextu nově vytvořeného nástroje ARCLib jsou zcela nové a pro dosažení cíle logické ochrany digitálních dokumentů nezbytné. Postupy jsou zcela nové i s ohledem na využité dílčí softwarové nástroje. Poprvé v ČR byl definován model odvození finančních a personálních nákladů dlouhodobé ochrany. Jedinečná je i vnitřní struktura informačních balíčků, jež byla v rámci projektu vytvořena. Metodika obsahuje též doporučení dalších nutných postupů v rámci dosažení dlouhodobé ochrany. Uživatelům softwarového řešení ARCLib poskytuje doporučení, jak výše zmíněné postupy aplikovat, jak provádět správu dat v systému a hodnotit rizika. Zároveň popisuje způsob uložení dat, jejich identifikaci a strukturu.

Popis uplatnění metodiky

Metodika bude uplatněna zejména u uživatelů systému ARCLib, ale umožňuje i nezávislé využití. Tato metodika je rozdělena do dvou hlavních částí, které jsou spolu provázány, ale přesto mohou do určité míry fungovat samostatně. Z tohoto rozdělení vyplývá i dvojí určení metodiky. V první (obecnější) rovině je metodika pro logickou ochranu digitálních dokumentů určena všem institucím a jejich odborným pracovníkům, kteří mají dlouhodobé uchovávání těchto dokumentů na starosti. Popisuje principy péče o dlouhodobé uchovávání, nutné požadavky na dlouhodobá úložiště, pravidla důvěryhodných repozitářů a nároky na provozující instituce. V tomto ohledu má metodika ambici být základním dokumentem v oboru v ČR a poskytovat pracovníkům paměťových institucí (nejen knihoven) metodickou podporu při plánování a správě systémů dlouhodobé ochrany digitálních dokumentů bez ohledu na konkrétní technické řešení nebo vymezení na jednotlivé druhy dokumentů. Význam této činnosti je již nyní značný a do budoucna se bude ještě zvyšovat.

Z hlediska samotné metodiky je však důležitější druhá část speciálně určená pro uživatele systému ARCLib. Jak již bylo výše řečeno, dlouhodobé uchovávání digitálních dokumentů neznamená jen využívání určitého softwaru případně ověřeného hardwaru. Ty představují jen nutné prostředky pro péči o uložené dokumenty. Jádro úkolů spojených s logickou ochranou digitálních dokumentů spočívá v dodržování zdokumentovaných procesů, v užívání osvědčené praxe a v kvalifikovaném personálu. Konkrétní část metodiky definuje postupy, jak provádět úkony logické ochrany v systému ARCLib. Popisuje jednotlivé funkcionality systému, jeho mapování na funkční prvky OAIS, architekturu informačních balíčků, úkoly, které je nutné vykonávat z hlediska vytvoření důvěryhodného systému, a navrhuje základní strukturu dokumentace nutné pro potvrzení důvěryhodnosti. Metodika tedy představuje podrobnou dokumentaci, která definuje postupy jakými dosáhnout úspěšné aplikace principů logické ochrany na digitální dokumenty uložené v systému ARCLib. Postupy v každém ze systémů jsou částečně odlišné a nelze je jednoduše přebírat. Metodika proto tvoří nedílnou součást užívání systému ARCLib, který nabízí odlišný způsob péče o uložené dokumenty než komerční systémy. Vzhledem k open source charakteru celého systému je nutné, aby jeho uživatelé měli k dispozici veřejně dokumentovanou metodiku. Lze navíc očekávat, že uživateli systému budou zejména krajské knihovny a specializované knihovny, zejména pro vědecká a výzkumná data. U knihoven těchto typů nelze očekávat, že by měly k dispozici plné spektrum odborníků, kteří by byli schopni sami definovat všechny nezbytné procesy dlouhodobého uchovávání v kontextu jejich sbírky. Právě tato funkčně vymezená komunita (tj. kurátoři repozitářů) by měla metodiku využívat při správě svých dat.

Systém ARCLib a postupy doporučené metodikou budou v rámci projektu "ARCLib – komplexní řešení pro dlouhodobou archivaci digitálních (knihovních) sbírek" vyžity v Knihovně Akademie věd ČR, v. v. i. Vzhledem k veřejnému určení metodiky není třeba uzavírat další smlouvy o jejím využívání.

Metodika bude po certifikaci Ministerstvem kultury umístěna do Národního úložiště šedé literatury NUŠL a volně zpřístupněna. Autory metodiky i jejího popisu použitého v tomto článku jsou členové řešitelského týmu projektu ARCLib Jan Hutař, Andrea Miranda, Eliška Pavlásková, Zdeněk Vašek a Zdeněk Hruška.

Citace odborné literatury:

[1] MELICHAR, Marek; HUTAŘ, Jan. České paměťové instituce a digitální data – historický exkurz, současný stav a předpokládaný vývoj I. Duha [online]. 2013, roč. 27, č. 4 [cit. 2017-10-03]. Dostupný z WWW: <http://duha.mzk.cz/clanky/ceske-pametove-instituce-digitalni-data-historicky-exkurz-soucasny-stav-predpokladany-vyvoj>. ISSN 1804-4255.

[2] LTP SAFE. LTP SAFE [online]. 2009 [cit. 2017-10-03]. Dostupné z: http://www.aipsafe.cz/cs/for-public-sector/ltp

[3] Systémy pro přenos dat a informací z kosmického prostoru – Audit a certifikace důvěryhodných digitálních úložišť: Space data and information transfer systems – Audit and certification of trustworthy digital repositories = Systèmes de transfert des informations et données spatiales – Audit et certification des référentiels numériques de confiance : ČSN ISO 16363 : schváleno v září 2011 ve Washingtonu, DC, USA. 1. vyd. Praha: Úřad pro technickou normalizaci, metrologii a státní zkušebnictví, 2014, 53 s.

[4] MELICHAR, Marek; HUTAŘ, Jan. České paměťové instituce a digitální data – historický exkurz, současný stav a předpokládaný vývoj III.. Duha [online]. 2014, roč. 28, č. 2 [cit. 2015-04-22]. Dostupný z WWW: <http://duha.mzk.cz/clanky/ceske-pametove-instituce-digitalni-data-historicky-exkurz-soucasny-stav-predpokladany-vyvoj-1>. ISSN 1804-4255.

[5] UML Activity Diagrams. Archivematica [online]. 2011 [cit. 2017-10-03]. Dostupné z: https://www.archivematica.org/wiki/UML_Activity_Diagrams

Prevádzka informačného systému CAIR a portálu Slovakiana

Peter Selecký, Národné osvetové centrum

Abstrakt
V rámci národných projektov Centrálna aplikačná infraštruktúra a registratúra a Harmonizácia informačných systémov bol vytvorený komplexný informačný systém pozostávajúci z viacerých zložiek aktuálne prevádzkovaný Národným osvetovým centrom. Cieľom príspevku je prostredníctvom zamerania sa na jednotlivé moduly vymedziť základnú funkcionalitu informačného systému CAIR. Nevyhnutnou súčasťou funkcionality je pripojenie jednotlivých pamäťových a fondových inštitúcií, ktoré vzhľadom na rozsah pripojenia využívajú rôzne kategórie služieb. V rámci prevádzky dochádza k postupnej optimalizácii využívania zdrojov vrátane poskytovaných služieb. Súčasťou informačného systému ja aj následná prezentačná vrstva pozostávajúcu z portálu kultúrneho dedičstva Slovakiana.

ÚVOD

Národné osvetové centrum bolo prijímateľom národných projektov Centrálna aplikačná infraštruktúra a registratúra (CAIR) a Harmonizácia informačných systémov (HIS), ktoré sa od ostatných projektov Operačného programu Informatizácia spoločnosti, prioritnej osi 2 „Rozvoj pamäťových a fondových inštitúcií a obnova ich národnej infraštruktúry“ (OPIS PO2) líšili tým, že nerealizovali samotnú digitalizáciu kultúrneho dedičstva, ale vytvárali také podmienky pre pamäťové a fondové inštitúcie, aby bolo možné na centrálnej úrovni nakladať so vzniknutým digitálnym obsahom.

Projekt Centrálna aplikačná infraštruktúra a registratúra (CAIR) budoval systém na centrálnej úrovni pre všetky inštitúcie v rezorte kultúry. V rámci tohto systému vznikol národný portál, národné registre – databázy údajov o objektoch kultúrneho dedičstva

a systém na prezentovanie kultúrnych a digitálnych objektov pre odbornú aj laickú verejnosť.

Projekt Harmonizácia informačných systémov (HIS) mal za úlohu „harmonizovať" digitalizujúce inštitúcie – tzn. zabezpečiť ich softvérovú a hardvérovú úroveň tak, aby sa mohli napojiť na systém vybudovaný v projekte CAIR, pracovať s dátami a zasielať ich na archiváciu.

CENTRÁLNA APLIKAČNÁ INFRAŠTRUKTÚRA A REGISTRATÚRA (CAIR)

Cieľom projektu bolo vytvorenie centrálneho systému rezortu kultúry, ktorý zabezpečí evidenciu a prezentáciu kultúrneho dedičstva Slovenska. V prepojení s Centrálnym dátovým archívom (CDA) vznikla unikátna platforma pre pamäťové a fondové inštitúcie Slovenska na sprístupňovanie, evidovanie a dlhodobé archivovanie kultúrneho dedičstva Slovenska v digitálnej podobe. Súčasťou tejto platformy je novovybudovaný systém národných registrov (register kultúrnych objektov, register digitálnych objektov, register autorít, register digitalizácie, register autorských práv), v ktorom sú zaradené všetky digitalizované objekty z digitalizačných projektov jednotlivých pamäťových a fondových inštitúcií (PFI) na základe kritérií nových jednotných metodík pre všetky zapojené subjekty rezortu kultúry. V neposlednom rade je potrebné spomenúť aj prezentačnú vrstvu, ktorú aktuálne tvorí portál kultúrneho dedičstva Slovakiana, prostredníctvom ktorého sa verejnosť môže dostať k digitálnemu kultúrnemu dedičstvu Slovenskej republiky.

Odborným pracovníkom v inštitúciách sa vďaka centrálnej evidencií dát zjednodušilo vyhľadávanie údajov o autoroch, autorských právach, kultúrnych a digitálnych objektoch. Interné katalogizačné systémy v PFI sa naplno harmonizovali na infraštruktúru CAIR, aby sa záznamy automaticky dopĺňali prírastkovým harvestingom. Prostredníctvom kooperácie PFI, integrácie dát do centrálneho dátového modelu a do piatich národných registrov sa zjednotilo sémantické vyhľadávanie a obohacovanie údajov na odbornej medzi sektorovej úrovni. Integráciou informačných systémov PFI do národných registrov, ktoré sú vzájomne previazané sa vytvorila najrozsiahlejšia centrálna infraštruktúra v rezorte MK SR.

Národný register digitalizácie (NRD) slúži pre centrálny informačný systém na evidenciu priebehu procesu digitalizácie kultúrnych objektov. Projekt digitalizácie je základnou plánovacou jednotkou, v rámci ktorej sa definujú vybrané kvantitatívne a kva-

litatívne parametre procesu digitalizácie. Primárnou úlohou národného registra digitalizácie je evidencia projektov digitalizácie, ich stavov a zložiek, vrátane evidencie stavov spracovania objektov, evidencie digitalizačných pracovísk a evidencie vybraných profilov PFI[1] v kontexte projektu CAIR.

Národný register digitálnych objektov (NRDO) je jedným z národných registrov, ktorý spravuje digitálne objekty evidované v CAIR. Poskytuje priestor pre evidenciu údajov o digitálnych objektoch, ich registráciu v globálnej schéme urn:nbn:sk:[2] a modifikáciu vybraných údajov digitálneho objektu. NRDO neposkytuje priamo priestor pre uloženie digitálneho obsahu PFI. Digitálne objekty sú uložené v samostatnom module prístupné len úzkej skupine používateľov, aby sa zachovali dôležité prezentačné deriváty. Používatelia systému NRDO prezerajú obsah digitálnych objektov, majú možnosť doplniť popisné údaje a štruktúrne mapy za účelom zvýšenia kvality zobrazovaných metadát na portáli Slovakiana a zároveň v NRDO majú možnosť nastaviť, ktoré záznamy budú publikované na portáli Slovakiana.

Národný register kultúrnych objektov eviduje a identifikuje kultúrne objekty spravované rôznymi PFI v Slovenskej republike za účelom zabezpečenia interoperability zdrojov údajov, koordinácie, vykazovania priebehu procesu digitalizácie a poskytovania základných informácií o kultúrnych objektoch. Jeho úlohou nie je komplexné uchovávanie a spravovanie metadát o objektoch kultúrneho dedičstva, ale evidencia poskytnutých údajov zo zdrojových inštitúcií a ich obohacovanie o vybrané údaje.

Národný register autorských práv (NRAP) eviduje v kontexte ochrany a evidencie práv duševného vlastníctva údaje týkajúce sa autorských práv a práv súvisiacich s autorskými právami na úrovni kultúrnych objektov, digitálnych objektov a prípadne častí jednotlivých objektov. Na základe týchto informácií sa vyhodnotí možnosť sprístupnenia objektov na portáli Slovakiana vrátane dovoleného spôsobu zobrazenia.

Národný register autorít (NRA) bude poskytovať odbornej verejnosti informácie o autoritách ako referenčných entitách používaných v katalogizačnom procese v odvetví kultúry, v rámci národnej infraštruktúry vybudovanej na základe projektu CAIR. Harmonizované záznamy o autoritách budú zároveň sprístupňované aj laickej verejnosti prostredníctvom portálu Slovakiana.

1 Profily definujú spracovanie a riadenie dát a metadát v CAIR, ktoré prichádzajú od inštitúcií.

2 Register URN:NBN: spravuje v súlade s RFC 3188 centrálna národná registračná autorita príslušnej krajiny. Funkciu centrálnej registračnej autority v Slovenskej republike vykonáva Slovenská národná knižnica. Správu registra URN:NBN zabezpečuje Slovenská národná knižnica spolupráci s Univerzitnou knižnicou v Bratislave.

Portál Slovakiana je novovybudovaná prezentačná platforma, ktorej cieľom je sprístupniť výsledky digitalizácie kultúrneho dedičstva Slovenska širokej verejnosti. Webový portál primárne umožňuje prístup na jednom mieste ku kultúrnym objektom v jednoduchej podobe, používateľsky prívetivejšom prostredí v porovnaní s národnými registrami, ktoré sú určené odbornej verejnosti.

Slovakiana prináša viac ako milión kultúrnych objektov v digitálnej podobe cez elektronické služby a funkcionalitu využiteľnú laickou a odbornou verejnosťou. Práve vďaka vybudovanej infraštruktúre ponúka Slovakiana aj rozšírené možnosti vyhľadávania, prezerania a sťahovania digitálnych objektov vo vysokom rozlíšení s možnosťou následného využitia v prípade, ak to umožňuje aktuálny stav právnej ochrany. Na portáli Slovakiana si návštevník vďaka funkciám personalizácie obsahu môže zo sprístupnených objektov vytvárať vlastné zbierky podľa svojich kritérií, diskutovať o nich s verejnosťou aj odborníkmi, zdieľať obsah, podávať návrhy na aktualizáciu alebo doplnenie obsahu, sledovať a byť informovaný o témach, ktoré ho zaujímajú a pod. Všetky objekty nesú aj informáciu o ich dostupnosti z pohľadu súvisiacej právnej ochrany. Vďaka týmto funkcionalitám by sa Slovakiana postupom času mala stať nástrojom využívaným v oblasti e-learningu, ale napríklad aj v oblasti kreatívneho priemyslu. Slovakiana sa zároveň stáva národným agregátorom pre celoeurópsky portál kultúrneho dedičstva Europeana.

HARMONIZÁCIA INFORMAČNÝCH SYSTÉMOV (HIS)

Hlavným cieľom projektu bolo aktualizovať, rozšíriť a skvalitniť informačné systémy pamäťových a fondových inštitúcií, dobudovať infraštruktúru pamäťových a fondových inštitúcií na národnej úrovni a zabezpečiť interoperabilitu s národným projektom CAIR.

Okrem uvedeného patrili medzi ciele aktualizácia a rozšírenie existujúcej hardvérovej infraštruktúry PFI, na ktorej prevádzkujú svoje informačné systémy, vybudovanie lokálneho technického vybavenia pre potreby realizácie a podpory konverzie zo strany PFI a ostatného postprocesingu, úprava informačných systémov PFI tak, aby mohli spolu s centrálnymi systémami CAIR, CDA a informačnými systémami digitalizačných pracovísk byť ich integrovanou súčasťou.

Výstupy projektu majú mnohostranné využitie. Pamäťové a fondové inštitúcie môžu využiť výsledky projektu realizáciou činností, ako napríklad:

- sprístupňovanie, prezentácia a popularizácia vlastného fondu prostredníctvom

moderných informačno-komunikačných kanálov, (využitie integračných rozhraní a spoločných rezortných modulov),
- efektívne procesy evidencie, katalogizácie, digitalizácie, spracovania, prezentácie a dlhodobej ochrany obsahu prostredníctvom podpory centrálnych riešení.

Rezort kultúry môže v rámci budovania konkurencieschopného a inovačného prostredia v oblasti kultúry vecne profitovať z tvorby nových druhov medzisektorových služieb nad celým kultúrnym dedičstvom, ktoré doteraz nebolo z dôvodu chýbajúceho centrálneho prepojenia možné poskytovať.

V rámci harmonizácie informačných systémov boli vytvorené nasledovné moduly:

Editor SIP profilov (ESP)

V rámci priebehu digitalizácie kultúrneho dedičstva bolo nevyhnutnosťou zaistiť transmisiu výsledkov digitalizácie jednak do centrálneho dátového archívu (CDA) z dôvodov dlhodobej archívnej ochrany a do CAIR z dôvodu prezentácie, sprístupneniu a prípadnému ďalšiemu spracovaniu.

Na splnenie úspešného prenosu do CDA a CAIR bolo nevyhnutnosťou vytvoriť zdrojové informačné balíky vo formáte SIP. Formát SIP a proces jeho vloženia určujú cieľové systémy CDA a CAIR. Skutočnosťou ostáva, že SIP balík určený pre CDA sa môže diferencovať od SIP balíka, ktorý je určený pre systém CAIR a to predovšetkým svojím obsahom. Takáto diferencovanosť nastáva aj napriek skutočnosti, že formát popisu a štruktúra obsahu vychádza z rovnakých noriem.

Na samotné vyprodukovanie SIP balíkov, ich validáciu a export do CDA a CAIR, je možné na strane PFI použiť existujúci modul na expedíciu (MES). Avšak aj napriek tejto skutočnosti môže byť samotná konfigurácia modulu MES pre konkrétne výstupy digitalizácie netriviálna a záleží od komplikovanosti štruktúry a formátov vytvárajúcich obsah. Táto konfigurácia je v module pre expedíciu (MES) zachytená do podoby digitalizačného profilu, ktorý:
- popisuje štruktúru PSP (balík dát, ktorý producent posiela na spracovanie)
- vymedzuje pravidlá pre validáciu vstupu
- vymedzuje pravidlo pre vytvorenie identifikátora SIP
- vymedzuje pravidlo pre podpisovanie obsahu SIP
- vymedzuje premenu metadát z daného PSP do SIP
- vymedzuje pravidlá pre mapovanie obsahu z daného PSP do SIP

Modul pre expedíciu SIP (MES) je distribuovaný prostredníctvom niekoľkých digitalizačných profilov (dProfilov) a zabezpečuje centrálnu registráciu, uloženie a samotné dodávanie dProfilov.

Predmetom modulu ESP je na jednej strane vytvorenie, modifikácia a správa dProfilov pre koncového používateľa, na strane druhej samotné vytvorenie SIP balíka pomocou interpretácie dProfilu.

ESP modul umožňuje vytvorenie balíkov SIP, či už prostredníctvom modulu pre expedíciu SIP (MES) alebo samostatne. Používateľovi dáva možnosť zadefinovať nový, prípadne zmeniť existujúci profil pre vytváranie SIP balíkov, určených pre CAIR alebo CDA.

Modul ESP disponuje:

- editorom štruktúry vstupných údajov (graficky resp. formou vyplnenia formulára umožní popísať štruktúru adresára a dát, ktoré reprezentujú výstup digitalizácie (PSP)
- editorom validačných pravidiel kontrolujúcich vstupné údaje (PSP)
- možnosťou exportu šablóny do formátu dProfilu určený pre expedičný modul (MES)
- sprievodcom (Wizard) vytvorením nového profilu
- Integráciou na Katalóg dProfilov v CAIR; oprávnený používateľ môže použiť, modifikovať alebo vytvárať nové profily, ktoré je možné zdieľať medzi viacerými prevádzkami resp. digitalizačnými pracoviskami. Katalóg profilov obsahuje (predpripravenú) sadu základných profilov
- Generátorom SIP, ktorý dokáže dávkovo vytvárať SIP balíky na základe vstupných PSP integráciou vybraného dProfilu

Editor digitálneho objektu – METS editor (EDO-METS)

V priebehu digitalizácie kultúrneho dedičstva vznikali a vznikajú digitálne objekty na rôznych úrovniach, s rôzne bohatým obsahom a v rôznej kvalite. Isté formy digitálnych objektov vznikajú už v digitalizačných pracoviskách ako PSP (voľný formát, ktorý zvyčajne definuje jeho tvorca – digitalizátor), neskôr sú digitálne objekty zaznamenané v podobe SIP (vždy vo formáte METS) pre účely expedície do CDA alebo CAIR, následne vo finálnej podobe v repozitári CAIR (taktiež vo formáte METS) resp. sú archivované v CDA.

Z popisovaného procesu je zrejmé, že formát METS je primárnym nástrojom pre popis digitálneho objektu, ktorého cieľom je transportovanie alebo exportovanie (prípadne

importovanie) digitálneho objektu medzi rôznymi aktérmi a systémami. Zároveň ostáva v platnosti, že METS ako XML formát určený na popis digitálneho objektu je rozsiahly a bohatý a môže disponovať rôznymi ďalšími formátmi, kdeže METS ako taký je v neposlednom rade aj kontajnerom pre vklad špecializovaných informácií ako sú napr. popisné metadáta vo formáte MODS, DublinCore alebo MARCXML, administratívne metadáta vo formáte MIX alebo XMP atď.

Editor digitálneho objektu – METS editor je aplikácia bežiaca lokálne, ktorá umožňuje používateľovi (autorovi resp. správcovi digitálneho objektu – DO) vytvoriť resp. modifikovať aj komplexný digitálny objekt lokálne, (t.j. v jeho prostredí) zapísaný v štandarde METS. Týmto spôsobom môže používateľ dosiahnuť maximálnu kvalitu (popisu) digitálneho objektu v zmysle kritérií CAIR aj CDA. Zároveň je táto aplikácia vhodná na prípravu jednotlivého DO pre vklad do CDA alebo podanie do CAIR, opravu resp. manuálne spracovanie DO, vytvorenie odvodeného DO z objektu vybraného z archívu a pod.

Generický modul pre riadenie integrácie v digitalizácii (GRID)

Generický modul pre riadenie integrácie v digitalizácií procesne manažuje postup a samotnú integráciu s registrami CAIR a lokálnymi systémami PFI pri príprave a priebehu digitalizácie, priebeh samotnej digitalizácie, ale predovšetkým integráciu pri prenose výstupov do CAIR a integráciu pri archivácii do CDA.

Medzi základne funkcionality modulu GRID patrí:

- off-line import dát z evidenčného systému (vo formáte CSV prípadne v XML) alebo integrácia s evidenciou kultúrnych objektov v registroch CAIR – importujú sa nutné polia pre identifikáciu kultúrnych objektov,
- manuálne pridávanie nových kultúrnych objektov,
- rozdelenie projektov podľa digitalizačných pracovísk,
- administrácia kont so zaradením používateľov do rolí (kurátor, konzervátor, operátor skenera, manažér a pod.),
- vizualizácia stavu projektu – súčty/podiely kultúrnych objektov podľa: stavu spracovania, typu kultúrnych objektov,
- integrácia na reporting v CAIR a integrácia za účelom získania stavu spracovania a archivácie z CAIR a CDA.

Príprava digitalizácie:
Modul pokrýva úvodnú časť digitalizačného workflow, kde dochádza ku ohodnoteniu, ošetreniu a prípadnému zásahu (konzervovanie, oprava a pod), a to:

- filtrácia podľa projektu, typu KO, umiestnenia, typu zásahu,
- vyhľadávanie objektov podľa identifikátora alebo názvu objektu,
- možnosť meniť atribúty v procese hodnotenia,
- evidencia jednotlivých krokov zásahu (zaznamenanie odkedy, dokedy, kto a aký zásah robil),
- funkčnosť modulu končí stavom KO „pripravený na digitalizáciu“ po vykonaní zásahov (ak boli potrebné).

Modul pre digitalizáciu:

- vyhľadávanie objektov podľa identifikátora (lokálneho alebo NID z NRKO) alebo názvu objektu pre potreby digitalizátora,
- filtrácia podľa projektu, typu objektu, stavu digitalizácie a časového hľadiska,
- workflow digitalizácie,
- zaradenie objektu do spracovania (končí presunom spracovaných objektov na úložisko),
- informatívne údaje (počty zdigitalizovaných objektov za rôzne časové obdobia),
- reporty ohľadom naplánovaných objektov a ich vlastnosti (typ, zásah, stav objektu) za jednotlivé projekty,
- zoznam objektov so všetkými atribútmi + náhľadový obrázok a možnosť filtrovania podľa týchto atribútov,
- odkaz na lokálne uloženie výsledkov digitalizácie v prípade, ak je umiestnenie výstupu formalizované na základe atribútov KO,
- obohatenie atribútov KO o vybrané metadáta z digitalizácie (napríklad identifikácia skenera) pre potreby štatistického modulu,
- evidencia transportu objektov samostatne a v transportných balíkoch,
- integrácia s EDO-METS pre manuálnu úpravu štruktúry digitálneho objektu a zvýšenie kvality jeho popisu pred archiváciu a prezentáciou,
- príprava na integráciu s CAIR a CDA – identifikácia, určenie kódov projektov a kódov balíkov registrovaných v CAIR, určenie spracovateľských profilov (v CAIR) a digitalizačných profilov (v CDA).

Modul štatistík:

- možnosť výberu podľa kritérií (projekt, typ, digitalizátor, zariadenie),
- filtrácia podľa časového hľadiska,
- tabuľkové aj grafické výstupy,
- možnosť tlače resp. export do formátov pdf, word a excel,
- štatistiky pre modul konzervovania,
- výstupné tabuľky pre fakturáciu (výpočty podľa typu objektu, miesta snímania a vykonaného zásahu),

- integrácia a reporting do NRD v CAIR – stavy a počty KO v spracovaní.

Integrácia:
V procese je možné použiť (pred-) pripravené akcie a konektory pre technickú interakciu procesu s ďalšími systémami, ako sú:

- rozhranie na služby Národného registra kultúrnych objektov (NRKO) pre výber KO určených pre digitalizáciu,
- rozhranie na služby Národného registra digitalizácie (NRD) pre plánovanie a reporting stavu digitalizácie,
- rozhranie na MES pre vklad do CDA,
- rozhranie na MES pre podanie obsahu do CAIR,
- rozhranie pre získavanie stavu archivácie objektu v CDA,
- rozhranie pre získavanie stavu spracovania a prezentácie objektu v CAIR.

Záver

V rámci projektov Centrálna aplikačná infraštruktúra a registratúra (CAIR) a Harmonizácia informačných systémov (HIS) bol vybudovaný komplexný informačný systém, ktorý dokáže plniť viacero úloh rezortu kultúry v oblasti digitalizácie kultúrneho dedičstva. Samotná digitalizácia je náročným procesom, ktorý musí postupovať podľa vopred definovaných štandardov tak, aby bolo možné zabezpečiť jednak dlhodobú ochranu a archiváciu výstupov digitalizácie, ale aj ich následné použitie.

Editor SIP profilov (ESP), Editor digitálneho objektu – METS editor (EDO-METS) a Generický modul pre riadenie integrácie v digitalizácii (GRID) sú moduly prispievajúce k zjednodušeniu niektorých krokov, ktorú sú nevyhnutnou súčasťou postupu digitalizácie.

Pre potreby evidencie boli vytvorené Národné registre prevádzkované Národným osvetovým centrom, ktoré poskytujú komplexný prehľad o kultúrnom dedičstve Slovenskej republiky a aktuálnom stave digitalizácie. Vytvorený systém je priebežne aktualizovaný a dopĺňaný, tak aby dokázal zohľadniť meniace sa prostredie a požiadavky pamäťových a fondových inštitúcií. Celý systém a súvisiace aktivity v rámci digitalizácie je nevyhnutné vnímať aj z pohľadu sprístupňovania kultúrneho dedičstva verejnosti. Na tento účel slúži centrálny portál kultúrneho dedičstva Slovakiana, ktorý prierezovo zobrazuje kultúrne a digitálne objekty rôznych pamäťových a fondových inštitúcií.

Standardizace při tvorbě digitálních dokumentů jako základ digitální archivace

Ladislav Cubr, Národní knihovna České republiky

Abstrakt
Široce přijímanou tezí digitální archivace (*digital preservation*) je, že rizika dlouhodobého uchovávání digitálních dokumentů, kterým musí čelit digitální repozitáře, lze významně snížit tehdy, pokud je již sama produkce dokumentů řízena s ohledem na jejich trvalou udržitelnost. Tento článek se souhrnně zabývá otázkami standardizace produkce z hlediska potřeb digitální archivace v kontextu uchovávání fondů digitálního dědictví. Rozebírá hlavní oblasti, kterým je třeba věnovat pozornost při plánování a vytváření digitálních dokumentů v paměťových institucích, a zaměřuje se především na jeden z hlavních typů digitálních dokumentů – digitalizáty fyzických dokumentů.

1 Úvod

Základem standardizace v oblasti digitální archivace je referenční rámec OAIS. Tento rámec, popsaný v normě ISO 14721, se zaměřuje na otázky dlouhodobého uchovávání a zpřístupňování digitálních dokumentů v digitálním repozitáři (1). Norma však do určité míry pokrývá také otázku produkce, a to v konceptu dohody o dodávání dat mezi vkladatelem a repozitářem a s ní související specifikací balíčku SIP. V praxi často dochází k tomu, že repozitář nemá možnost tuto specifikaci ovlivnit způsobem, který by odpovídal požadavkům digitální archivace. Pak je při převodu do balíčku AIP nutné vykonávat formátovou normalizaci či jiné náročné úpravy, které mohou být v některých případech obtížně realizovatelné. V tomto článku se zabýváme optimálním případem, kdy archiv a vkladatel (digitalizující útvar) plně spolupracují na standardizaci produkce s ohledem na požadavky digitální archivace, tj. dlouhodobého uchovávání a zpřístupňování digitálních dokumentů uživatelům navzdory technologickým změnám.

2 Stanovení podoby dokumentu

Základním požadavkem na úspěšné řízení životního cyklu digitálního dokumentu je, aby ještě před zahájením digitalizace byla přesně stanovena podoba dokumentu, který má být cílem digitalizace. To může působit jako samozřejmost, ale z hlediska požadavků digitální archivace jde o poměrně náročný úkol. Předem vymezit je potřeba nejen základní intelektuální entitu (ve smyslu standardu PREMIS), kterou bude digitální dokument představovat, ale také jeho klíčové vlastnosti. U digitalizátů je navíc nutné specifikovat stupeň věrnosti předloze.

Specifikace základní intelektuální entity znamená určení typu dokumentu a míry granularity. Tato základní entita pak bude představovat dokument uložený v balíčku SIP, který bude předmětem bibliografického popisu i zpřístupňování uživatelům v prezentačním systému.[1] To vyžaduje, aby předloha byla ještě před zahájením její digitalizace kvalitně zkatalogizována (resp. aby byl její záznam pečlivě zkontrolován, upraven nebo rozšířen). Ve specifických případech může být popis doplněn až při vytváření metadat pro balíček SIP.[2]

Klíčovými vlastnostmi zde rozumíme vlastnosti, kterými musí disponovat výsledný digitální dokument (uložený v balíčku SIP) a které musejí být zachovány napříč všemi procesy odvozování (transformace a kopírování) v repozitáři a při jakémkoliv aktu reprodukování dokumentu počítačovou technologií (zejména softwarem, např. při zobrazení stránek v digitální knihovně). Klíčové vlastnosti jsou zde chápány v širším významu než pojem signifikantní vlastnosti (*significant properties*)[3] standardu PREMIS (2, s. 50) – jsou vztaženy i na fázi produkce. Potřeba stanovit klíčové vlastnosti reflektuje skutečnost, že v dlouhodobém horizontu je prakticky vyloučeno zachovat digitální objekty v jejich původní podobě, a to zejména z důvodu zastarávání formátů.[4] Tyto vlastnosti musejí být stanoveny před zahájením produkce – jako cíl digitalizace, který bude zároveň cílem následného uchovávání a zpřístupňování digitalizátu v současnosti i budoucnosti. Producent dat tedy musí vzít v potaz, jaké vlastnosti dokumentu lze dlouhodobě zachovat, a tomu přizpůsobit specifikaci dokumentu pro daný projekt. Nemá smysl vytvářet data s příliš složitými funkcemi, pokud existuje vysoké riziko, že

1 Základní intelektuální entitou může být například gramodeska, mapa, kniha (včetně vícesvazkové knihy), svazek knihy, číslo periodika apod.

2 Typicky v případě, kdy je základní intelektuální entita zvolena ve vyšší míře granularity (např. článek periodika), než kterou popisuje katalogizační záznam.

3 V normě ISO 14721:2012 je obdobný pojem „transformační vlastnost informací“ (transformational information property) (1).

4 Blíže k problematice formátů viz loňský konferenční příspěvek (10).

je nebude možné dlouhodobě uchovávat. Klíčové vlastnosti by měly být zaznamenány do projektové dokumentace, která by měla být uchovávána společně s digitalizáty v repozitáři i po skončení digitalizace. Dokumentace bude sloužit jednak jako pravidlo pro budoucí migrace (co musí být zachováno), jednak jako explicitní deklarace toho, co mohou uživatelé od zpřístupňovaných dokumentů očekávat a vyžadovat. Klíčové vlastnosti digitalizátů je vhodné rozdělit do dvou skupin: vlastnosti převáděné digitalizací (např. barevné vlastnosti knihy) a vlastnosti přidané při zpracování, které jsou nativně digitální. Nativně digitální vlastnosti znamenají obohacení dokumentu o funkce, které fyzická předloha nikdy neměla – u digitalizovaných knih jde například o plnotextovou prohledatelnost. První skupina vlastností pak odkazuje na otázky věrnosti digitalizace a očekávání s ní spojená v širší uživatelské komunitě.

Specifikace věrnosti znamená určit, k jaké úrovni abstrakce a k jakému časovému hledisku předlohy se digitalizační projekt vztahuje. K popisu úrovně abstrakce lze užít model FRBR.[5] Principem většiny současných projektů masové digitalizace v knihovnách je věrnost obrazovým a strukturálním vlastnostem exempláře tištěné knihy v jejím současném stavu (tj. se známkami stárnutí). Filmové archivy se při digitalizaci mohou řídit jiným principem, založeným na rekonstrukci původního stavu předlohy (filmového pásu) v době vzniku; v takovém případě je nezbytné odborné digitální restaurování v průběhu zpracování původních skenů.[6] Je však také zcela legitimní zdigitalizovat knihu tak, aby byl převeden pouze text v jeho struktuře (tj. bez stránkování a dalších znaků vydání), pokud je v projektu věrnost specifikována na úrovni vyjádření (podle FRBR). Tímto způsobem již proběhla řada digitalizací, například na Oxfordské univerzitě.[7] Základní podmínkou tedy je, aby specifikace věrnosti byla zaznamenána, uchovávána a zpřístupňována v projektové dokumentaci. Pro tuto specifikaci lze čerpat z řady existujících směrnic a standardů. Pro běžnou digitalizaci knih (věrnost na úrovni exempláře) lze užít směrnici DLF,[8] která uvádí výčet parametrů, které musejí věrné digitalizáty splňovat (zachování úplnosti a vzhledu původních stránek včetně tonality a barvy, původní posloupnost stránek ad.) (3). Textový formát TEI[9] je vhodným způsobem, jak standardizovat digitalizaci na úrovni vyjádření.

5 Model FRBR stanovuje tyto čtyři úrovně abstrakce: dílo (work), vyjádření (expression), manifestace (manifestation) a jednotka (unit) (14). V případě knih představuje manifestaci vydání knihy a jednotku exemplář knihy v rámci jednoho vydání.

6 Viz např. současný digitalizační projekt českého Národního filmového archivu (12).

7 Viz např. sbírka Shakespearových her (http://ota.ox.ac.uk/desc/3014).

8 Digital Library Federation (Federace digitálních knihoven)

9 http://www.tei-c.org/

3 Specifikace objektu CDO v balíčku SIP

Hlavním předmětem činnosti repozitáře je zachovat informační obsah (*content information*), resp. jeho klíčové vlastnosti, a to napříč technologickými změnami. Informační obsah se skládá z objektu CDO (*content data object*) a interpretačních informací (*representation information*). Jeden objekt CDO může být v závislosti na implementaci tvořen jedním nebo více soubory a tyto mohou být v jednom nebo více formátech; podle toho můžeme rozdělit objekt CDO na odlišné komponenty. Například digitální film je obvykle objektem CDO tvořeným třemi komponentami: zvukovou (zvukový kodek), obrazovou (video kodek) a strukturální (kontejnerový formát).[10] Informační obsah lze dále odlišovat jako uložený a reprodukovaný. Uložený informační obsah je uchováván v balíčku AIP; skládá se z objektu CDO a interpretačních informací, které mohou být uloženy ve dvou podobách: jako technická metadata (informace o formátu apod.) nebo odkaz na dokumentaci uloženou také v repozitáři.[11] Z hlediska uživatelů je důležitý především reprodukovaný informační obsah, který představuje intelektuální entitu vnímatelnou uživatelem (např. knihu zobrazenou v digitální knihovně). V tomto případě jsou interpretační informace tvořeny samotným zpřístupňovacím softwarem. Interpretační informace jsou rozhodným prvkem digitální archivace. Musejí být vždy dostupné – formát objektu CDO musí být dobře popsán a na základě tohoto popisu musí být možné nalézt software k jeho adekvátní reprodukci. Z hlediska zpřístupňování to znamená, že softwarové řešení prezentačního systému musí být navrhováno a vyvíjeno v závislosti na vývoji internetových prohlížečů, požadavcích čtenářů a změnách zpřístupňovacího softwaru (grafických prohlížečů apod.). Repozitář musí vybírat a testovat takové nástroje, které dokáží informační obsah reprodukovat bezchybně a v souladu s jeho deklarovanými klíčovými vlastnostmi. Pokud podpora formátu upadá, znamená to, že interpretační informace přestávají být dostupné, což ohrožuje zachování informačního obsahu. Pak je nutno vytvořit (formátovou migrací v repozitáři) nový objekt CDO a shromáždit k němu nové interpretační informace.

Z výše uvedeného vyplývá, že vhodným postupem pro vytváření balíčku SIP je, aby objekt CDO v něm obsažený nemusel být v rámci následného převodu do balíčku AIP nikterak měněn.[12] To především předpokládá, aby formáty komponent objektu CDO

10 Viz např. loňský konferenční příspěvek o archivaci audiovizuálních děl (13, s. 60).

11 Například v podobě specifikace formátu JPEG 2000 v souboru v PDF.

12 Je myšlen balíček AIP první verze. Budoucí (novější) verze balíčku AIP mohou zahrnovat formátovou migraci původního objektu CDO.

byly archivačními formáty.[13] Optimální je také to, aby jeden balíček SIP obsahoval vždy právě jeden objekt CDO reprezentující jeden dokument.

Volba konkrétního archivačního formátu objektu CDO je závislá na stanovených klíčových vlastnostech dokumentu a dále na zvoleném implementačním modelu digitalizace. Běžný model je takový, že základním předmětem snímání je jedna stránka knihy a základním objektem uložení jeden soubor reprezentující tuto stránku. Lze však zvolit odlišné modely – například snímání knihy po dvojstránkách[14] nebo uložení celého zdigitalizovaného dokumentu do jediného souboru.[15] Pokud klíčové vlastnosti zahrnují plnotextovou prohledatelnost, pak užití rastrových formátů není dostačující – je potřeba připojit další, textovou komponentu (výstup z OCR). Pokud je věrnost chápána na úrovni vyjádření, může být celý objekt CDO tvořený jediným souborem ve formátu TEI XML.

Specifikace objektu CDO znamená určení digitální identity jeho komponent, v prvé řadě formátu. Pro běžnou digitalizaci tištěných monografií a periodik jde o tři komponenty: obrazovou (pro věrnou obrazovou reprodukci) ve formátech TIFF nebo JP2; textovou (pro plnotextovou prohledatelnost) ve formátu ALTO a strukturální (ta je nezbytná pro vyjádření vztahů mezi komponentami a jejich částmi, jednotlivými soubory) ve formátu METS.[16] Formát je však třeba specifikovat podrobněji: udat oficiální plný název formátu[17], jeho verzi, identifikátor PUID (pokud existuje) a formátový profil (specifické nastavení v rámci formátu). Součástí identity obrazové komponenty jsou navíc speciální obrazové vlastnosti, které jsou relativně nezávislé na formátu (např. prostorové rozlišení, bitová hloubka, barevný prostor nebo barevný profil). Pro jejich specifikaci lze doporučit digitalizační směrnice americké iniciativy FADGI[18] (4). Optimálním postupem je, aby u těch komponent, pro které je relevantní datová komprimace, byla použita bezeztrátová komprese (případně nekomprimovaný formát). Každá ztrátová komprese může z hlediska uchovávání znamenat problém, a to bez ohledu na to, zda je při reprodukci dokument vnímán jako bezeztrátový.[19]

13 Blíže viz loňský příspěvek věnovaný formátové strategii (10).

14 Viz projekt Octavo, který digitalizoval staré tisky jako otevřené dvojstránky (11, s. 41).

15 Viz například digitalizáty starších diplomových prací na Univerzitě Karlově, které byly snímány do rastrových formátů, ale po zpracování uloženy do formátu PDF. Jeden soubor v PDF obsahuje jednu diplomovou práci a takto je uchováván i zpřístupňován čtenářům.

16 Specificky jde o část formátu METS (záznam fyzické mapy). Tato část je součástí objektu CDO (nikoliv metadat), protože tvoří samu strukturu objektu CDO.

17 Například formáty ALTO a METS jsou podtypem formátu XML.

18 Federal Agencies Digital Guidelines Initiative

19 Viz případ tzv. „vizuálně bezeztrátové“ komprese užité v profilu formátu JP2, která je však matematicky ztrátová (9, s. 4).

Specifikace objektu CDO v balíčku SIP tedy v prvé řadě závisí na stanovených cílech digitalizace (klíčové vlastnosti) a volbě procesu produkce (implementační model). Teprve na základě těchto voleb lze provést konkrétní specifikaci objektu CDO a formátu a dalších vlastností jeho komponent, která musí být jednoznačná a precizní. Užití specializovaných mezinárodních směrnic pro produkci digitálních dokumentů může tuto specifikaci usnadnit.

4 Specifikace odvozování a nástrojů

Po volbě podoby dokumentu a specifikaci objektu CDO je nutné provést specifikaci celého postupu jeho odvozování. Ten začíná digitalizací fyzické předlohy, pokračuje zpracováním původních dat (ořezy obrazů, restaurování filmů, korekce šumu apod.) a končí finalizací objektu CDO pro balíček SIP. Celý proces odvozování musí být nejprve otestován, zdali je schůdný, v případě zjištění problémů upraven, poté zdokumentován v digitalizačním projektu a až nakonec zrealizován ve vlastní digitalizaci. Determinujícím prvkem odvozování jsou užitá digitalizační zařízení, softwarové aplikace (grafické, zvukové, videové editory apod.) a nástroje (kodeky apod.). Obecně se doporučuje, aby bezprostředním výstupem digitalizace byl nekomprimovaný formát. Digitalizační zařízení tedy musí tuto volbu umožňovat. Je také nutno zjistit přesný počet generací objektu CDO. V současné praxi můžeme nalézt tyto řetězce formátových konverzí v průběhu vytváření balíčku SIP: TIFF (tj. pro snímání i zpracování je užit tentýž formát, který je i konečným archivačním formátem); TIFF – JP2; RAW – TIFF; RAW – DNG – TIFF; RAW – TIFF – JP2; RAW – DNG – TIFF – JP2. Omezení vyplývající z dostupných nástrojů vedou například k tomu, že z formátu RAW nelze přímo vytvořit formát JP2 (proto musí být vytvořen meziformát TIFF), nebo že je při zpracování formátu RAW nutný převod do meziformátu (DNG).

Přesné stanovení řetězce odvozování a počtu generací objektu CDO je důležité jednak pro stanovení toho, co bude předmětem metadatového popisu (tj. kolik zaniklých generací objektu CDO bude popsáno), jednak pro posouzení toho, zda při odvozování nedošlo k významné ztrátě klíčových vlastností. Pokud je výsledným formátem bezeztrátově komprimovaný JP2 nebo nekomprimovaný TIFF, ale v průběhu zpracování byl jako meziformát užit například formát JPEG (nebo bylo do JPEG snímáno), pak je zjevné, že výsledná obrazová komponenta, ačkoliv jde formálně o bezeztrátovou kompresi, bude reálně obsahovat ztrátovou obrazovou informaci.

5 Specifikace validace a vytváření dodatečných informací

V procesu vytváření dokumentu by měl být v maximální míře využíván digitální otisk. Ten by měl být generován bezprostředně po vytvoření nového souboru[20] a zkontrolován po každém kopírování, resp. před každou následnou transformací. Po každé transformaci by měl být také zkontrolován formát (tj. zda je validní). Je možné, že původně vybraný nástroj nebude schopen vytvářet ze zdrojových dat validní formát, a proto musí být nahrazen jiným (tento problém řeší výše uvedené prvotní testování).

Stěžejní je samozřejmě komplexní validace balíčku SIP, která musí zahrnovat validaci jak všech komponent objektu CDO, tak metadat. Optimálně by měly být k dispozici vhodné formátové i datové validátory[21] a v digitalizační dokumentaci zapsáno, jaké konkrétní validátory (včetně verzí) byly užity, s případným uvedením jejich omezení (známé chyby apod.), resp. údaje o tom, že validace nemohla být provedena, protože validátor neexistuje (v tomto případě pak bude muset být validace vykonána později v repozitáři, až bude dostupný odpovídající validátor). Validace balíčku SIP by měla být vyváženě rozdělena mezi vkladatele (producenta dat) a repozitář a v dohodě o dodávání dat by měl být uveden postup pro řešení případných oprav. Konečný objekt CDO by měl validovat vkladatel i repozitář. Validace na straně vkladatele slouží jako kontrola kvality produkce, zatímco validace v repozitáři patří mezi nezbytné opatření digitální archivace. Předchozí generace objektu CDO, které se po vytvoření balíčku SIP smažou, však může validovat pouze vkladatel. Zde je cílem validace zvýšení jistoty, že v celém řetězci odvozování nedošlo například k softwarové chybě.

Balíček AIP musí kromě objektu CDO a technických metadat zaznamenávajících interpretační informace obsahovat též archivační informace (*preservation description information*), které jsou nezbytné pro procesy uchovávání informačního obsahu a doklady o jeho autenticitě. Interpretační informace musí vždy vytvářet repozitář. Může je také vytvářet producent balíčku SIP. Primární zdrojem by měl být proces charakterizace (extrakce metadat ze souboru). Repozitář by měl také uchovávat všechny užité standardy (formátové specifikace, metadatové standardy apod.), které jsou také typem interpretačních informací. Z archivačních informací, které musí vytvářet vkladatel, jsou stěžejními dvě skupiny: identifikační a provenienční informace. Identifikační informace zahrnují digitalizační dokumentaci (popis výše uvedených specifikací) a popis identity konkrétního dokumentu (bibliografický záznam, identifikátor předlohy na úrovni

20 Digitální otisk vytvořený později již může být otiskem porušeného souboru a je tedy zavádějící.

21 Validátory datového toku souboru, jako jsou obrazové validátory (např. ImageMagick).

exempláře a manifestace, údaj o stavu předlohy, perzistentní identifikátor digitalizátu, datum vytvoření digitalizátu aj.). Provenienční informace zahrnují popis procesu odvozování, validací a všech zaniklých generací objektu CDO. Tyto informace může archiv získat pouze obtížně, pokud vůbec, a do značné míry to platí i o identifikačních informacích. Většinu uvedených informací je vhodné zaznamenat v podobě metadat, která tvoří součást balíčku SIP. Současným standardem je serializace metadat do formátu XML a jejich uložení v samostatných souborech (5, s. 131). Existují dva typy metadatových standardů pro potřeby současné digitalizace: univerzální (METS, PREMIS, MODS) a dokumentově specifické. METS slouží primárně pro záznam informací o zabalení (*packaging information*) balíčků SIP (součástí této funkce je vnoření dalších metadatových schémat) a jako datový formát pro strukturální komponentu. PREMIS slouží pro zápis archivačních a obecných interpretačních informací, z hlediska produkce je důležitý pro záznam událostí (odvození a validace). V registru řízených slovníků udržovaných Kongresovou knihovnou jsou doporučené hodnoty pro některé metadatové elementy PREMIS, které by měly být používány.[22] Velmi důležitou letošní novinkou je vydání rozšířeného řízeného slovníku pro události v PREMIS (6). Specifické standardy (např. MIX pro rastrová obrazová data, documentMD pro textové dokumenty ad.) slouží zejména jako technická metadata (pro záznam interpretačních informací) pro specifické typy dat.

Metadatové standardy jsou nutně determinující prvek. Předepisují určitou logiku zápisu a jejich užití do značné míry umožňuje zbavit se nutnosti v každém projektu detailně zjišťovat, jaké všechny typy informací je potřebné shromažďovat. Řídit se mezinárodními standardy znamená také vyšší záruku interoperability. Zároveň to vyžaduje vyčkávat na případné doplnění elementů pro něco, co zatím popsat nelze, v budoucí nové verzi standardu či zcela novém standardu. Vytvářet vlastní metadatový standard je intelektuálně náročné a přináší riziko ztráty interoperability i nejistotu vývoje. Vytvářet externí schémata (jako rozšíření stávajících standardů) je smysluplnější. Metadatový profil užitý pro digitalizaci musí obsahovat informace o všech využitých metadatových standardech, jejich verzích a elementech z těchto standardů vybraných pro potřeby konkrétního digitalizačního projektu. Tento profil musí být samozřejmě v souladu s těmito standardy. Měl by však také obsahovat informace o užitých řízených slovnících významných organizací, případně také vlastní řízené slovníky pro hodnoty vybraných dalších elementů, které nejsou standardizovány těmito organizacemi. To zvyšuje možnosti kontroly a správy interpretačních a archivačních informací. Bibliografické údaje by měly být vždy získávány z katalogizačního systému (to vyžaduje kvalitní kontrolu záznamů, resp. rekatalogizaci), a teprve pak převáděny do popisného standardu pro digitální objekty (MODS). Interpretační informace by měly být v maxi-

22 http://id.loc.gov/

mální možné míře získávány přímo ze souborů užitím ověřených metadatových extraktorů (JHOVE, jpylyzer aj.). Postup získávání metadat by neměl být založen na údajích, které se přednastaví do digitalizačního systému (např. název skeneru pro jednu linku) a systém je následně jen automaticky přiděluje všem dokumentům dané linky. Pro objekty CDO tvořené obrazovými soubory je také vhodné využívat zabudovaná EXIF metadata (7). Iniciativa FADGI vydala doporučení, jaká minimální metadata EXIF je třeba extrahovat (8). Tento způsob získávání metadat je vzhledem k vysoké standardizaci EXIF možno označit za důvěryhodný způsob plnění metadat. Současně umožňuje předejít problému, kdy v lince dojde k náhradě snímacího zařízení, ale zapomene se tento údaj zadat do přednastavených hodnot. Výstupy extraktorů však nelze vždy jednoduše namapovat do mezinárodních metadatových standardů. Metadatový profil by tedy měl také obsahovat popis způsobu převodu informací z výstupů metadatových extraktorů do elementů metadatového profilu. Současně však může být vhodné (zejména ve fázi archivace) také uložit samotná metadatová schémata těchto extraktorů a doplnit je jako externí schémata.[23]

6 Úrovně identifikace

Identifikaci je třeba rozlišovat na několik úrovních. Nejnižší úrovní je systém pojmenovávání souborů při digitalizaci, pro který lze užít směrnici FADGI. V ní je doporučeno užití názvů, které jsou jedinečné v celém digitalizačním projektu (tedy nejen v rámci balíčku SIP jednotlivého digitalizátu), a zavedení znakových restrikcí (maximální počet znaků, užití malých písmen apod.) (4, s. 79-81). Pro identifikaci formátu je klíčový identifikátor PUID, který však nemusí být dostupný pro užitý typ formátu. Určitou možností je podílet se na vytvoření nového záznamu v registru PRONOM.

Nejdůležitější je perzistentní identifikace fyzické předlohy na relevantní úrovni FRBR (v případě knih na úrovni vydání), perzistentní identifikace digitálního dokumentu jakožto digitalizátu této předlohy a jejich vzájemné propojení (zejména v prezentačním a katalogizačním systému a v metadatech balíčků SIP a AIP). Pro perzistentní identifikaci digitálního dokumentu je vhodné užít takový identifikační systém, který splňuje následující požadavky: jednoznačnost (jednoznačně identifikuje dokument v daném kontextu); facilitace užití[24] (usnadnění práce s identifikátorem); perzistence (identifi-

23 Například výstupy nástroje FITS lze začlenit jako externí schéma do PREMIS (v rámci sekce „objectCharacteristicsExtension“).

24 Je vhodné řídit se jednoduchým ekonomickým principem: struktura identifikátoru musí být na jedné straně co nejkratší, aby se uživatelům s identifikátory co nejsnáze pracovalo, a so-

kátor, který byl v určitém okamžiku přidělen jednomu dokumentu, již nikdy nesmí být přidělen znovu žádnému jinému); globálnost (mezinárodní jedinečnost) a trvalé přesměrovávání (tj. nezávisle na URL adrese – užitím resolveru). Takovými systémy jsou zejména DOI, Handle a URN:NBN. Perzistentní identifikátor by měl být přidělován tak, aby dokázal odlišit různé digitalizáty téže předlohy (z hlediska odlišného vlastníka a digitalizačního projektu).

Perzistentní identifikátor digitálního dokumentu slouží v informačních balíčcích jako trvalý identifikátor informačního obsahu, který zůstává neměnný navzdory formátovým migracím objektu CDO, které jsou vykonávany v repozitáři nebo pro potřeby zpřístupnění. Měl by být viditelný uživatelům (v metadatech zobrazených spolu s dokumentem) a uživatel by měl být obeznámen s pravidly daného identifikačního systému, aby věděl, na jaké adrese nalezne resolver pro přesměrovávání na aktuální URL adresu dokumentu. Takto koncipovaný perzistentní identifikátor pak lze užít i v citační praxi. Identifikační systém bude trvale nabízet službu dereference (přesměrování nebo prezentace metadat popisujících dokument).[25]

7 Závěr

Pro standardizaci vytváření digitálních dokumentů je třeba využívat nejen mezinárodní metadatové standardy, osvědčené postupy pro výběr archivačních formátů a důvěryhodné identifikační systémy, ale také specializované směrnice pro produkci dokumentů určitého typu. Standardizace musí vycházet z určení základní intelektuální entity a klíčových vlastností dokumentu (včetně popisu přístupu k věrnosti digitalizace) a implementačního modelu užitého při digitalizaci. Tyto volby by měly být precizně specifikovány před zahájením digitalizace a zaznamenány v projektové dokumentaci, která by měla být následně dlouhodobě uchovávána a zpřístupňována uživatelům. Takováto dokumentace je pak základem pro budoucí formátové konverze i posuzování toho, nakolik repozitář dostává závazkům, které byly v digitalizačním projektu deklarovány.

učasně musí být schopna pojmout předpokládaný rozsah identifikovaných objektů. Dalším prvkem facilitace je užití omezené sady znaků (např. pouze alfanumerické znaky).

25 Blíže viz (16).

Citované zdroje

1. ČSN ISO 14721. *Systémy pro přenos dat a informací z kosmického prostoru – Otevřený archivační informační systém – Referenční model*. 2. vyd. Praha: Úřad pro technickou normalizaci, metrologii a státní zkušebnictví, 2014, 97 s. Třídící znak 31 9620.
2. PREMIS EDITORIAL COMMITTEE. *PREMIS Data Dictionary for Preservation Metadata* [online]. Version 3.0. Washington (DC): Library of Congress, June 2015, rev. Nov 2015, viii, 273 s. [cit. 2017-09-20]. Dostupné z: http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf.
3. THE DIGITAL LIBRARY FEDERATION BENCHMARK WORKING GROUP. *Benchmark for Faithful Digital Reproductions of Monographs and Serials* [online]. Version 1. Washington (DC): Digital Library Federation, December 2002, 6 s. [cit. 2017-09-20]. Dostupné z: http://old.diglib.org/standards/bmarkfin.pdf.
4. FEDERAL AGENCIES DIGITAL GUIDELINES INITIATIVE, Still Image Working Group. *Technical Guidelines for Digitizing Cultural Heritage Materials: Creation of Raster Image Master Files* [online]. Washington (DC): FADGI, Sep 2016, 100 s. [cit. 2017-09-20]. Dostupné z: http://www.digitizationguidelines.gov/guidelines/FADGI%20Federal%20%20Agencies%20Digital%20Guidelines%20Initiative-2016%20Final_rev1.pdf.
5. ZENG, Marcia Lei a QIN, Jian. *Metadata*. 2nd edition. Chicago: Neal-Schuman, an imprint of the American Library Association, 2016. xxvii, 555 stran. ISBN 978-1-55570-965-5.
6. *Preservation Events Controlled Vocabulary* [online]. Washington (DC): Library of Congress, 2017, Release date: 22 June 2017. [cit. 2017-09-20]. Dostupné z: http://www.loc.gov/standards/premis/v3/preservation-events.pdf.
7. CIPA DC-008-TRANSLATION-2012. *Exchangeable file digital still cameras: exif version 2.3* [online]. Tokyo: CIPA, 2012, 185 s. [cit. 2017-09-20]. Dostupné z: http://www.cipa.jp/std/documents/e/DC-008-2012_E.pdf.
8. FEDERAL AGENCIES DIGITIZATION INITIATIVE, Still Image Working Group. *Guidelines for TIFF Metadata: Recommended Elements and Format* [online]. Version 1.0. Washington (DC): FADGI, Feb 10, 2009, 5 s. [cit. 2017-09-20]. Dostupné z: http://www.digitizationguidelines.gov/guidelines/TIFF_Metadata_Final.pdf.
9. BUCKLEY, Robert. *JPEG 2000 as a Preservation and Access Format for the Wellcome Trust Digital Library* [online]. London: King's College London, Aug 2009, 17 s. [cit. 2017-09-20]. Dostupné z: http://wellcomelibrary.org/content/documents/22082/JPEG2000-preservation-format.pdf.

10. CUBR, Ladislav. Formátová strategie LTP úložiště NK ČR. In: *CDA 2016: formátové výzvy LTP: zborník príspevkov z 1. medzinárodnej konferencie o dlhodobej archivácii*. Bratislava: Univerzitná knižnica v Bratislave, 2016, s. 44-57. ISBN 978-80-89303-51-9. ISSN 2453-9406.
11. GRYCZ, Czeslaw Jan. Digitising Rare Books and Manuscripts. In: MACDONALD, Lindsay, ed. *Digital heritage: applying digital imaging to cultural heritage*. Amsterdam: Elsevier, 2006, s. 33-68. ISBN 0-75-066183-6.
12. *Digitální restaurování českého filmového dědictví* [online]. Praha: Národní filmový archiv, 2017 [cit. 2017-09-20]. Dostupné z: http://eea.nfa.cz/.
13. BUBESTINGER, Peter. File formats for audiovisual preservation : How to choose? In: *CDA 2016: formátové výzvy LTP: zborník príspevkov z 1. medzinárodnej konferencie o dlhodobej archivácii*. Bratislava: Univerzitná knižnica v Bratislave, 2016, s. 58-80. ISBN 978-80-89303-51-9. ISSN 2453-9406.
14. IFLA STUDY GROUP ON THE FUNCTIONAL REQUIREMENTS FOR BIBLIOGRAPHIC RECORDS. *Functional Requirements for Bibliographic Records: final report* [online]. Haag: IFLA, September 1997, as amended and corrected through Feb 2009, v, 137 s. [cit. 2017-09-20]. Dostupné z: https://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf.
15. METS: An Overview & Tutorial. *Metadata Encoding and Transmission Standard (METS)* [online]. Washington (DC): Library of Congress, February 9, 2016 [cit. 2017-09-20]. Dostupné z: http://www.loc.gov/standards/mets/METSOverview.v2.html.
16. CUBR, Ladislav et al. Srovnání vybraných národních identifikačních systémů užívajících identifikátory URN:NBN. *ProInflow: časopis pro informační vědy* [online]. 2016, **8**(1) [cit. 2017-09-20]. Dostupné z: http://www.phil.muni.cz/journals/index.php/proinflow/article/view/1220

Descriptive metadata for long-term preservation and the bibliographic management of digital surrogates

Szabolcs Dancs, National Széchényi Library, Hungary

Abstract
The role of national libraries as information providers is determined by the specific tasks they fulfil in the library system. Through digitizing cultural heritage they are obliged to provide access to a verified corpus of digitized objects which can serve as surrogates of original ones for any possible patron in the future. This obligation includes endeavouring to reflect content (text, illustrations, etc.) and also the typographical outfit in its totality. There is no need to emphasize that verification is a key concept when compiling our digitization policy. It is crucial (among others) from bibliographic aspect. The article intends to give an insight into the standard tools for verification and, in a more detailed way, into the management of descriptive metadata of digitized library materials in an RDA/MARC 21 environment.

The role of national libraries as information providers is determined by the specific tasks they fulfil in the library system. Through digitizing cultural heritage they are obliged to provide access to a verified corpus of digitized objects which can serve as surrogates of original ones for any possible patron in the future. This obligation includes endeavours to reflect content (text, illustrations, etc.) and also the typographical outfit in its totality. As it is laid down by James A. Jacobs and James R. Jacobs: “Books and other printed information packages (e.g., journals, newspapers) don’t just store and transport information; they also encode and present information. They *are* the user-interface: the layout of text on the page imparts meaning... The relative positions of text and non-text elements (position on the page, text and non-text adjacency) is itself

information that provides context and specifies inter-relationships between the two."[1] Of course, digitization could have other (secondary) aims from a national library perspective, for instance: building services upon digital corpus such as virtually reconstruction of collections or documents, etc.

Verification

Nevertheless there is no need to emphasize that verification is a key concept when compiling our digitization policy. It is crucial (among others) from bibliographic aspect. Until recently, most librarians haven't been concerned about putting any URL into the field '856' (Electronic Location and Access) of a MARC-based description of an analogue document without defining the relation between the electronic resource accessible via the URL and the original document described. Now we are supposed to identify entities and relations (between entities) according to FRBR and apply cataloguing rules (and systems) supporting the creation of entity records and the use of relation designators. If a digitized copy published via WWW can be considered as a publication (as it is suggested by the American AACR/RDA tradition), than it is possible to regard it as a Manifestation related to an Item (which latter is an example of the original, analogue document). The relation between the two might be defined as 'is a verified digital copy of'.

Figure 1

We also need to provide a link to the rules used for verification (included in the Digitization Policy of the organization concerned).

Figure 2

DSSOA and Digitization Policy

The *Digital-Surrogate Seal of Approval* (DSSOA) „is a tool for asserting that a digital surrogate accurately and completely replicates the content and presentation of a static, analog original such as a book". In other words "[a] user given sufficient access to such a digital surrogate would not need to consult the original".

DSSOA doesn't provide a detailed method for verification. The two major criteria a digital surrogate must meet according to the standard are the following:

> "1. **Completeness.** All pages of the original are fully and completely reproduced. No page or part of a page is obscured or blurred or masked or skipped or missing. There is a complete reproduction of each page in its entirety.
> 2. **Accuracy.** The original layout and appearance are preserved. All text is legible and at least as easily read as the original. There is no visual degradation as compared to the original. All text is clear and without any blur or distortion at the same size as the original. All images are clear and of a quality (e.g., color, resolution) equal to the original."

The rules for applying the DSSOA are specified in the following way:

> "The DSSOA may be applied to a bibliographically-identified item when:
> 1. The responsible organization has verified that both DSSOA criteria are met for that item.
> 2. The responsible organization provides a Statement of Verification.
>
> The Statement of Verification must:
> 1. Specify and describe the methodology used to verify compliance.
> 2. Confirm 100% compliance."

All these imply that the procedure of verification must be detailed by the organization itself e.g. as a part of its Digitization Policy. There are two usable documents to mention here.

Technical Guidelines for Digitizing Cultural Heritage[2] by the Federal Agencies Digital Guidelines Initiatives (FADGI) involves evaluation parameters such as sampling frequency, tone response, white balance error, illuminance non-uniformity, colour Accuracy, colour channel mis-registration, etc. Minimal technical requirements are listed in tables broken down according to document types (e.g. bound volumes: rare and special materials; documents (unbound): manuscripts and other rare and special materials; oversize items: maps, posters, and other materials with challenging features that will benefit from high resolution reproduction).

Another exploitable document is the practical guidelines for digitization of Deutsche Forschungsgemeinschaft (German Research Foundation)[3]. Beyond technical requirements this document addresses metadata issues as well. Sustainable usability of metadata requires compliance with cataloguing standards such as RDA and reference models such as CIDOC-CRM or IFLA FRBR / FRBRoo. Descriptive metadata must be provided according to material-specific standards:

- Metadata Encoding and Transmission Standard / Metadata Object Description Schema (METS / MODS) for printed texts and archival material
- Metadata Encoding and Transmission Standard / Text Encoding Initiative (METS / TEI) for manuscripts[4]
- Lightweight Information Describing Objects (LIDO) for (unique) visual and three-dimensional objects

It is also mandatory to provide access to descriptive metadata via an Open Archives Initiative (OAI) interface. Users also can exploit DFG-Viewer which is the reference implementation for the digitization standards of the DFG (METS / MODS, METS / TEI, OAI-PMH).

Detailing the above documents is out of the scope of this article, nonetheless we should note that from some aspects both of them are proved to be useful when compiling a Digitization Policy, but, unfortunately, they do not provide us with an exact and full methodology for a DSSOA-compliant verification.

The role of descriptive and bibliographic metadata in LTP

As *Angela Dappert* and *Markus Enders* state in their study issued in Information Standards Quarterly[5]:

"[a] preservation policy specifies digital preservation goals to ensure that:

- digital content is within the physical control of the repository;
- *digital content can be uniquely and persistently identified and retrieved in the future;*
- *all information is available so that digital content can be understood by its designated user community;*
- significant characteristics of the digital assets are preserved even as data carriers or physical representations change;
- physical media are cared for;

- digital objects remain renderable or executable;
- digital objects remain whole and unimpaired and that it is clear how all the parts relate to each other; and
- digital objects are what they purport to be."

According to the two authors "all of these preservation functions depend on the availability of preservation metadata". I might add that as far as those highlighted by me are concerned bibliographic metadata has an undeniable contribution.

Open Archival Information System (OAIS) reference model is the basic standard for long-term preservation, but, as it is emphasized by Dappert and Enders, "it does not define which specific metadata should be collected or how it should be implemented in order to support preservation goals". Specific metadata needed for long-term preservation is grouped into four categories by the authors:

> **"Descriptive metadata:** Describes the intellectual entity through properties such as author and title, and supports discovery and delivery of digital content. It may also provide an historic context, by, for example, specifying which print-based material was the original source for a digital derivative (source provenance).
>
> **Structural metadata:** Captures physical structural relationships, such as which image is embedded within which website, as well as logical structural relationships, such as which page follows which in a digitized book.
>
> **Technical metadata for physical files:** Includes technical information that applies to any file type, such as information about the software and hardware on which the digital object can be rendered or executed, or checksums and digital signatures to ensure fixity and authenticity. It also includes content type-specific technical information, such as image width for an image or elapsed time for an audio file.
>
> **Administrative metadata:** Includes provenance information of who has cared for the digital object and what preservation actions have been performed on it, as well as rights and permission information that specifies, for example, access to the digital object, including which preservation actions are permissible."

As we can notice here, bibliographic data for source provenance (i.e. the original, analogue document) might be considered as the part of the descriptive data. However it is not necessarily so, as we shall see it below.

When it is about long-term preservation and OAIS there are to relevant – XML-based – standards to mention. While *PREMIS Data Dictionary for Preservation Metadata*, which is based on the OAIS reference model, focuses on administrative and technical data, METS (*Metadata Encoding and Transmission Standard*) serves as a tool for packaging information, i.e. packaging archived digital objects and their associated metadata.

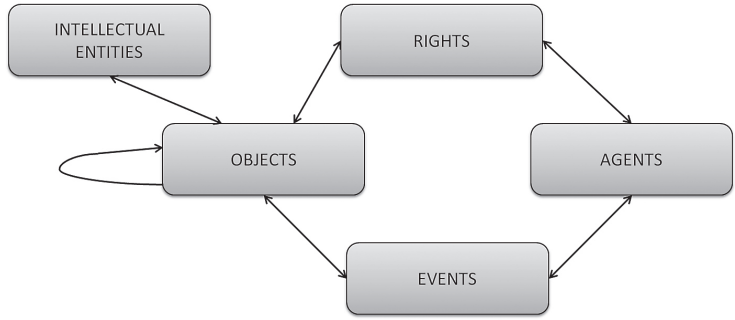


Figure 3

PREMIS is a general model consisting of *semantic units* considered to be important for digital preservation purposes. Semantic units are defined as properties of one of the five *entities* in the model (Intellectual Entities, Object, Event, Agent, Rights), they might be represented in a hierarchic way[6]:

Agent

3.1 agentIdentifier (M, R)
  3.1.1 agentIdentifierType (M, NR)
  3.1.2 agentIdentifierValue (M, NR)
3.2 agentName (O, R)
3.3 agentType (O, NR)
3.4 agentVersion (O, NR)
3.5 agentNote (O, R)
3.6 agentExtension (O, R)
3.7 linkingEventIdentifier (O, R)
  3.7.1 linkingEventIdentifierType (M, NR)
  3.7.2 linkingEventIdentifierValue (M, NR)
3.8 linkingRightsStatementIdentifier (O, R)
  3.8.1 linkingRightsStatementIdentifierType (M, NR)

3.8.2 linkingRightsStatementIdentifierValue (M, NR)
3.9 linkingEnvironmentIdentifier (O, R)
3.9.1 linkingEnvironmentIdentifierType (M, NR)
3.9.2 linkingEnvironmentIdentifierValue (M, NR)
3.9.3 linkingEnvironmentRole (O, R)

Although PREMIS is based on OAIS, as mentioned before, "[its] scope does not overlap precisely with the metadata requirements described in the reference model – for example, the OAIS concept of Descriptive Information is not covered by PREMIS"[7]. Meanwhile bibliographic information might function as a major source for PREMIS data as it is well-shown by the data overlap between the MARC21 Bibliographic Format and the PREMIS Data Dictionary.[8]

As for METS, it provides us with a schema to encode OAIS archival information packages (Submission Information Package, Archival Information Package, Dissemination Information Package). However there are other schemas in the arena such as BagIt or Fedora Object XML (FOXML), METS is a widely implemented one which is generally associated to the OAIS model. Its seven major sections are:

- **METS Header** <metsHdr> – metadata describing the METS document, e.g. information on the creator
- **Descriptive Metadata** <dmdSec> – might contain (1) a pointer to external metadata (<mdRef>), e.g. to a bibliographic record in MARC format, (2) internally embedded metadata (<mdWrap>), or (3) both (see below)
- **Administrative Metadata** <amdSec> – information on the "history" of the file: technical metadata, rights, *metadata on the original (analogue) source object*, provenance (master/derivative file relationships, migration, conversion, etc.)
- **File Section** <fileSec> – list of files constituting the digital object
- **Structural Map** <structMap> – the hierarchical structure for the digital object
- **Structural Links** <structLink> – a list of links between the components of the structural map
- **Behavior** <behaviorSec> – list of executable behaviors with content in the METS object

In External Descriptive Metadata the mdRef element provides an identifier (URN, URL, PURL, HANDLE, DOI) for retrieving metadata describing the digital object:

```
<dmdSec ID=»dmd01»>
<mdRef LOCTYPE=»URN» MIMETYPE=»application/xml» MDTYPE=»DC»
```

```
LABEL=» Kraus, Milan [et al.]: A kétfejű
macska»>urn:nbn:hu-5955</mdRef>
</dmdSec>
```

In this case the type of the retrieved data was DC but other values (e.g. MARC, MODS, EAD, DC) also might be defined.

For Internal Descriptive Metadata an mdWrap element is used:

```
<dmdSec ID=»dmd002»>
<mdWrap MIMETYPE=»text/xml» MDTYPE=»DC» LABEL=»Dublin Core
Metadata»>
   <xmlData>
     <dc:title>A kétfejű macska</dc:title>
     <dc:creator>Milan Kraus</dc:creator>
     <dc:creator>Ján Ondruš</dc:creator>
     <dc:creator>Štefan Strážay</dc:creator>
     <dc:publisher>NAP Kiadó</dc:publisher>
     <dc:type>text</dc:type>
    </xmlData>
   </mdWrap>
    </dmdSec>
```

(ID attribute of <dmdSec> is used in the structural map to link a particular division of the document hierarchy to a particular <dmdSec> element.)

Descriptive Metadata contains (or refers to) the description of the digitized object which is – in the terms of FRBR/RDA – *a Manifestation related to an Item* (the original object). The Relationship Designator could be 'is a verified digital copy of' (in case of verification process undertaken). The entity record on Item is embedded (or referred) as a part of the Administrative Data.

Figure 4

Source Metadata <sourceMD> element of Administrative Data is used for recording information on the original source:

```
    <sourceMD ID=»source01»>
      <mdRef LOCTYPE=»URL» MIMETYPE=»application/xml»
MDTYPE=»MARC»
      LABEL=» Kraus, Milan [et al.]: A kétfejű macska»>http://
www.mokka.hu/web/guest/record/-/record/MOKKAS0002300863</
mdRef>
    </sourceMD>
```

Description of digital surrogates in RDA and MARC 21

As it was already mentioned, a digitized copy published via WWW can be considered as a published reproduction of the original, analogue document. According to RDA you ought to use the following guideline when describing reproductions (RDA 1.11):

"When describing a facsimile or reproduction, record the data relating to the facsimile or reproduction in the appropriate element. Record any data relating to the original manifestation as an element of a related work or related manifestation, as applicable."

For the sake of philological exactness it is also necessary to record the relationship between the original Item and the digital version. There is also a possibility for that in RDA. Basic instructions on recording relationship to related Item are listed in 28.1.1. The section includes the following example as well (28.1.1.3):

"*Reproduction of:* ADM 55/40"

In MARC 21 field '776 – Additional Physical Form Entry' is defined for this purpose:

776 08**$i**electronic reproduction of (item):**$a**Szerb Antal (1901-1945)**$**tA Pendragon-legenda**$d**Budapest : Franklin, 1934**$h**280 p. ; 18 cm**$o**170.020**$w**000002965227

The subfields we used here were the following:

subfield codes	subfield names
$a	Main entry heading (NR)
$d	Place, publisher, and date of publication (NR)
$h	Physical description (NR)
$i	Relationship information (R)
$n	Note (R)
$o	Other item identifier (R)
$t	Title (NR)
$w	Record control number (R)

Relationship information (subfield 'i') is expressed by Relationship Designators which can be found in Appendix 'J' in RDA. In this case we used the following one (J.5.2):

> "electronic reproduction of (item): An item in an analog format that is transferred to a digital format"

We identified the related Item using subfield 'o' (Other item identifier) however we could also make a short note ('Call number of the original item: 170.020'), or, in case of an *Item record* is at disposal, we could use subfield 'w' (Record control number). (Here we recorded the record control number of the Manifestation concerned.)

For a full MARC 21 record about a digital surrogate accessible via WWW the following data also needs to be recorded as a minimum requirement:

- Content Type (336)
- Media Type (337)
- Carrier Type (338)
- Physical Medium (340) – eventually, to record generation by inserting values such as 'original', 'master', 'derivative master' (see RDA 3.10.1.3)
- Digital File Characteristics (347)
- Electronic Location and Access (856)

Let's see an example for recording relevant information as part of a MARC 21 record on a digital surrogate:

336 ##**$a**text**$2**rdacontent
337 ##**$a**computer**$2**rdamedia
338 ##**$a**online resource**$2**rdacarrier

347 ##**$a**text file**$b**PDF**$2**rda
776 08**$i**electronic reproduction of (item):**$a**Szerb Antal (1901-1945)**$**tA Pendragon-legenda**$d**Budapest : Franklin, 1934**$h**280 p. ; 18 cm**$n**Call number of the original: 170.020**$w**000002965227
856 40 **$3**OSZK – Digitális Könyvtár**$u**http://nbn.urn.hu/N2L?urn:nbn:hu-136665

Virtual Items compiled by combination of different digitized parts of various Items can be described by using subfield 'g' (Related parts) in the repeatable field '776':

776 08**$i**electronic reproduction of (item):**$a**Szerb Antal (1901-1945)**$**tA Pendragon-legenda**$d**Budapest : Franklin, 1934**$h**280 p. ; 18 cm**$o**170.020**$w**000002965227**$g**p. 1-72.

776 08**$i**electronic reproduction of (item):**$a**Szerb Antal (1901-1945)**$**tA Pendragon-legenda**$d**Budapest : Franklin, 1934**$h**280 p. ; 18 cm**$o**PTEVISz72**$w** MOKKAZ0004940112**$g**p. 72-280.

Conclusion

It might be concluded that RDA provides us with a tool to clearly express relation between entities such as of digital copy (considered to be a Manifestation) and original (analogue) Item however it currently misses some elements to transform outlined theory into practice, these are:

- A Relationship Designator 'verified electronic reproduction of (item):'
- A subfield in '776' to include information about the resource serving as the basement of verification ('Verified by')

RDA is a flexible standard and national versions might differ from the original, mainly by including additional instructions, therefore the suggested method can be implemented without any problem. We believe that philological exactness, which is expected from a national library, can be assured this way. We are also of the opinion that through active participation in European RDA Interest Group (EURIG) we shall find ways to discuss issues that are to emerge during the implementation period and later on.

Notes

1. http://dlib.org/dlib/march13/jacobs/03jacobs.html (accessed 18 September 2017)
2. Federal Agencies Digital Guidelines Initiative (FADGI), Still Image Working Group: Technical Guidelines for the Still Image Digitization of Cultural Heritage Materials. Approved by Working Group, September 2016 – http://www.digitizationguidelines.gov/guidelines/FADGI%20Federal%20%20Agencies%20Digital%20Guidelines%20Initiative-2016%20Final_rev1.pdf (accessed 19 September 2017)
3. DFG-Praxisregeln: „Digitalisierung“ / Deutsche Forschungsgemeinschaft – http://www.dfg.de/formulare/12_151/12_151_de.pdf (accessed 19 September 2017)
4. TEI, which is used e.g. in Manuscriptorium Project, is rather useful for philological process of digital items. In NSL we prefer to describe manuscripts according to RDA/MARC21.
5. Dappert, A. – Enders, M. (2010), “Digital preservation: metadata standards”, *Information Standards Quarterly*, Vol. 2 Issue 2, pp. 4-13., available at: https://www.loc.gov/standards/premis/FE_Dappert_Enders_MetadataStds_isqv22no2.pdf (accessed 19 September 2017)
6. http://www.loc.gov/standards/premis/v3/premis-hierarchical-3-0.html (accessed 19 September 2017)
7. Lavoie, B. (2014): The Open Archival Information System (OAIS) Reference Model: introductory guide (2nd Edition), available at: http://www.dpconline.org/docs/technology-watch-reports/1359-dpctw14-02/file (accessed 19 September 2017)
8. https://www.loc.gov/marc/bibliographic/mapping/MARC21vsPREMIStable.html (accessed 19 September 2017)

Projekt PREFORMA a aplikácia výsledkov projektu PREFORMA v Centrálnom dátovom archíve UKB

Juraj Strnisko, Univerzitná knižnica v Bratislave

Abstrakt
Cieľom projektu PREFORMA je zaviesť kvalitné štandardizované formáty pre uchovanie dát z pohľadu dlhodobej archivácie LTP (Long Term Preservation). S tým úzko súvisí použitie vhodných nástrojov na overenie (validáciu) formátov pred vložením do archívu. Tento príspevok opisuje vývoj tejto problematiky a načrtáva aplikáciu výsledkov projektu PREFORMA v Centrálnom dátovom archíve.

PREFORMA project and application of its result in the Central Data Archive in University Library of Bratislava

Ing. Juraj Strnisko, University library in Bratislava

Abstract:
Aim of the PREFORMA project is to address the challenge of implementing good quality standardised file formats for preserving data content in the long term preservation. This is closely related to the use of appropriate tools to validate formats prior to archiving. This article describes the development of this issue and outlines the application of the PREFORMA project results in the Central Data Archive.

1. Prečo PREFORMA

Problematika formátov dát z pohľadu dlhobobej ochrany LTP (Long Term Preservation) je veľmi obšírna. CDA ako LTP archív venuje tejto problematike systematickú pozornosť.

Vo svete existuje niekoľko projektov, ktoré sú zamerané na sledovanie vývoja jednotlivých formátov. CDA [1] ako východisko používa výsledky projektu PRONOM. PRONOM je vlastne akousi databázou všetkých existujúcich formátov. No z pohľadu LTP nie sú všetky formáty uvedené v PRONOME vhodné na archiváciu. V princípe ide o to, ako vybrať formáty, ktoré budú použiteľné aj o povedzme 20 či 50 rokov. To znamená, že sa budú dať prečítať alebo pretransformovať na iný v budúcnosti masovo podporovaný formát. Medzi takéto formáty nepatria tzv. propietárne formáty (napr. .doc, .xlsx, .mp3 a pod.)

Projekt PREFORMA vznikol ako reakcia na problémy komunity zaoberajúcej sa dlhodobým uskladnením dát.

2. Čo je PREFORMA

Názov PREFORMA vznikol skátením PREservation FORMATs [2]. Ide o projekt spolufinancovaný z fondov EU, ktorý oficiálne začal v roku 2014 a koncom roku 2017 skončí. Cieľom projektu je riešiť problém zavedenia kvalitných štandardizovaných formátov súborov na zachovanie dátového obsahu z dlhodobého hľadiska (LTP).

Hlavným cieľom je poskytnúť pamäťovým inštitúciám plnú kontrolu nad procesom testov zhody súborov voči štandardov pre dané súbory (tzv. validácia). Ide o kontrolu pri prípave dát do SIP (Submission Information Package) balíčkov, ktoré sa plánujú vkladať do archívov. Preto sa v rámci tohto projektu vyvinuli nástroje na kontrolu formátov súborov pred vkladom (tzv. validátory)

Jedná sa o tieto open-source nástroje:

VeraPDF na validáciu súborov PDF/A, [3]

DPFmanager na validáciu súborov TIFF, [4]

MediaConch na validáciu súborov MKV / FFV1 / LPCM, [5]

Výhodou týchto nástrojov je, že umožnujú kontrolovať formáty pri širokej škále nastavení. Sú k dispozícii aj na online testovanie (na web stránkach jednotlivých nástrojov). Ale taktiež sú voľne stiahnuteľné pre rôzne systémové platformy (Linux, Mac, Windows). Nástroje je možné parametrizovať. Výstup je možné konfigurovať. V prípade, že výsledok kontroly formátu je "Neúspešný", vedia tieto nástroje určité chyby aj opraviť. Vychádza sa zo štandardov pre jednotlivé typy formátov.

Okrem toho, sa ľudia združení v konzorciu PREFORMA podieľajú aj na aktualizácii formátových štandardov (ti-a.org, návrh štandardu TI/A, ktorý principiálne vychádza z TIFF/A).

Konzorcium PREFORMA tvorí 14 partnerov z 9 krajín Európy. Švédsky národný archív Riksarkivet je hlavným koordinátorom projektu. 9 partneri sú pamaťové inštitúcie (napr. Packed VZW z Belgicka, Beeld En Geluid – Holanský inštitút pre audiovíziu). Títo prevažne vytvárajú požiadavky na projekt, pripravujú dáta na testy. A potom je tu ešte 5 partnerov, ktorí majú na starosti realizáciu projektu s ich špecifickou expertízou. Sem patria 2 univerzity (zo Švédska a Talianska), inštitút Fraunhofer IDMT z Nemecka a 2 súkromné firmy, ktoré poskytujú expertízu v oblasti štandardov a komunikácie.

CDA sa formálne zaradila do PREFORMA komunity ešte v roku 2016 a má status externého člena. Na jeseň 2016 sa pracovníci CDA prvýkrát zúčastnili PREFORMA workshopu, ktorý sa konal v Berlíne. Okrem výmeny skúseností z oblasti LTP je možné na takýchto stretnutiach aj konzultovať nastavenie a riešiť problémy priamo s tvorcami (vývojármi) týchto validačných nástrojov.

3. Vklad do CDA (Ingest)

Vzhľadom k tomu, že CDA je implementované v súlade s OAIS STN ISO 14721:2014 [6] a STN ISO 27001:2013 pozostáva celý proces vkladu dát do archívu z viac ako 40 krokov. Tieto procesy možno pre zjednodušenie rozdeliť do 3 skupín:

Kontrola – profilu, certifikátov, formátov a antivírusová kontrola

Transformácia SIP na archívny informačný balíček AIP (Archived Information Package)

Uloženie AIP v 3 kópiach ("online" v lokalite vkladu, vytvorením tzv. synchronizačnej kópie pre druhú lokalitu a offline kópia, ktorá putuje do trezoru)

4. Formátová validácia CDA

V súčastnom období sa v CDA na validáciu používa identifikátor formátov Droid, využívajúci PRONOM databázu. Ten zistí (identifikuje) formáty v rámci SIP balíčka. Na overenie, či dané typy formátov v rámci balíčka sú správne, sa používa formátový validátor. Najčastejšie [7] sa v CDA používa validátor formátov Jhove (aktuálne vo verzii 1.7.2), napísaný v jazyku Java.

5. Aplikácia validátora vyvinutého v rámci PREFORMA

CDA registruje PFI (Pamäťové fondové inštitúcie), ktoré by výhľadovo chceli vkladať video súbory do archívu. V rámci spoločných analýz spolu s dodávateľom SW riešenia sa ako najvhodnejší video formát javí použitie formátu .MKV (tzv. Matroska). Preto v rámci posledného update-u aplikačného SW ktorý používa CDA, došlo aj k implementácii nástroja MediaConch a k úprave konfigurácie na validáciu videa. Odladenie zatiaľ prebehlo na 1 testovacom balíku. Žiadne z PFI doteraz nedodalo SIP balíky s MKV videom, na ktorých by bolo možné vykonať záťažové testy nového validátora.

Paralelne prebieha analýza a interné testy ďalších validátorov (veraPDF a DPFmanager). V súčasnosti je totiž možné do CDA vkladať PDF len do verzie 1.6 vrátane. PDF v najnovšej verzii 1.7 nie je zatiaľ podporované. Taktiež ani súbory vo formáte PDF/A.

6. Záver

Validácia súborov a výber vhodného formátu pre dlhodobé uchovanie dát sú témy, ktoré sú z pohľadu CDA a digitálnych archívov stále aktuálne. V CDA je už aktuálne nasadený nástroj MediaConch na validáciu video formátov MKV (s kodekom ffv1). Doteraz však nebol podrobený záťažovému testu kvôli nedostatku SIP balíkov obsahujúcich .MKV súbory.

Taktiež sa výhľadovo uvažuje o implementácii ostatných 2 validátorov z dielne PREFORMA, (VeraPDF a DPFmanager) príp. o nasadení vyššej verzie Jhove (ver. 1.16 vydaná v marci 2017 by už mala do určitej miery podporovať aj kontrolu/validáciu PDF vo verzii 1.7)

Literatúra

[1] Centrálny dátový archív. Dostupné z: http://cda.kultury.sk .

[2] Preforma project. Dostupné z: http://www.preforma-project.eu/ .

[3] VeraPDF. Dostupné z: http://verapdf.org/ .

[4] DPFmanager. Dostupné z: http://dpfmanager.org/ .

[5] MediaConch. Dostupné z: https://mediaarea.net/MediaConch/ .

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

[7] Rakús, Milan. Centrálny dátový archív a formátová stratégia CDA. In: Zborník CDA 2016, Formátové výzvy LTP. ISSN 2453-9406.

Standardizace Národní digitální knihovny

Zdeněk Vašek, Národní knihovna ČR

Abstrakt
Příspěvek naváže na předchozí vystoupení týkající se standardizace při digitalizaci a představí konkrétní principy Standardů Národní digitální knihovny, seznámí posluchače s využitými mezinárodními standardy, jejich vzájemnou provázaností a pravidly tvorby SIP balíčků. Dále se zaměří na význam Standardu v rámci dlouhodobého uchovávání, v čem spočívá jeho důležitost a proč je důležité, aby zpracovatelům omezoval možnosti využívání různých formátů dat a metadat. Podrobněji popíše omezení a výhody pojetí Standardu NDK a seznámí posluchače s konkrétními příklady."

Jedním z nezbytných opatření pro dlouhodobé uchovávání digitálních dokumentů je jejich standardizace, čímž je myšlen proces tvorby digitálních dokumentů, který je řádně popsán, děje se deklarovanými nástroji a sama vytvořená data odpovídají svému typu, respektive své datové specifikaci. Stejně tak musí být jednotný popis těchto dat. Jen pokud jsou tato pravidla dodržena, lze provádět efektivní správu dat, vyhledávat a řadit v rámci repozitářových systémů a provádět opatření na zajištění dlouhodobého uchovávání, zejména formátové migrace. Bez dodržení pravidel standardizace nelze zajistit hlavní cíl dlouhodobého uchovávání, tak jak ho definovala norma ČSN ISO 14721, tedy zachování informačního obsahu.[1] Bez přesné znalosti datového formátu nelze provádět složitější operace (zejména zmíněnou formátovou migraci) a může být ohroženo i pochopení obsahu, který při nedodržení pravidel pro popis nemusí být srozumitelný budoucím softwarům nebo uživatelům. Ve stručnosti zde byly shrnuty důvody o významu standardizace v dlouhodobém uchovávání digitálních dokumentů. V následující části se pokusíme tyto obecné představy ilustrovat na konkrétním případu Standardu Národní digitální knihovny.

1 *Systémy pro přenos dat a informací z kosmického prostoru – Otevřený archivační informační systém – Referenční model = Space data and information transfer systems – Open archival information system (OAIS) – Reference model = Systèmes de transfert des informations et données spatiales – Système ouvert d'archivage d'information (SOAI) – Modèle de référence: ČSN ISO 14721: schváleno v červnu 2012 ve Washingtonu, DC, USA*. Druhé vydání. Praha: Úřad pro technickou normalizaci, metrologii a státní zkušebnictví, 2014. 97 stran.

Projekt Národní digitální knihovna (respektive plným názvem „Vytvoření Národní digitální knihovny“) podpořený financemi z Integrovaného operačního programu EU proběhl v letech 2009-2014 a měl za cíl zdigitalizovat, dlouhodobě uložit a zpřístupnit 26 milionů stran knih a periodik ze sbírek Národní knihovny ČR a Moravské zemské knihovny. Aktuálně projekt pokračuje ve fázi udržitelnosti. Součástí aktivit spojených s projektem bylo i vytvoření Standardu NDK, který definuje metadatové schéma, povolené formáty, pravidla pro identifikátory a pravidla popisu určující, jak vytvářet ty části metadatového popisu, které nelze přebrat z katalogu, ale musí být vytvořena odbornými pracovníky digitalizační linky.[2] Standard prochází postupnou evolucí, jeho verze mají odlišné autory, ale kontinuita v názvech a verzování je dodržena. Hlavní účel Standardu je standardizovat produkci digitálních dokumentů vznikajících v knihovnách ČR. Standard předepisuje podobu vstupujících SIP balíčků, jejich strukturu a povolené metadatové elementy, případně i hodnoty těchto elementů.[3]

Za významný úspěch je možno považovat fakt, že Standard dosáhl širokého rozšíření a je využíván mnoha subjekty.[4] Prakticky veškerá probíhající digitalizace v knihovnách ČR se řídí tímto předpisem (knihovny digitalizující v podprogramu VISK 7 povinně, ostatní dobrovolně).[5] To přináší výhody jak v oblasti produkce, tak i následného zpřístupnění a dlouhodobého uložení. Vedle finančních úspor jde právě o účinnější a snáze poskytované služby dlouhodobého uchovávání. Podstatou celého Standardu je na jedné straně omezení formátů a struktury dat a metadat tak, aby do dlouhodobého úložiště nevstupovalo nic, co není popsané a o čem kurátoři nemají informace. Na druhé straně je Standard návodem pro ty, kteří se vzhledem k šíři odborných témat nemohou orientovat ve všech náležitostech spojených s produkcí a uchováváním digitálních dokumentů. Standard je v odborné gesci Oddělení pro standardy v rámci Odboru digitálních fondů

2 Standard NDK je dostupný z adresy http://www.ndk.cz/standardy-digitalizace (citováno 2017-09-28 – toto datum citace platí i pro všechny další URL odkazy, pokud nebude zaznamenáno jinak). Skládá se z definic metadatových formátů pro každý z typů dokumentů (např. LODROVÁ, Iveta a Jaroslav KVASNICA. *Definice metadatových formátů pro digitalizaci periodik: Dokument verze 1. 6* [online]. 2015 [cit. 2017-10-01]. Dostupné z: http://www.ndk.cz/standardy-digitalizace/DMFperiodika_16.pdf – Autoři původního dokumentu, ze kterého současná verze vychází, jsou Mgr. Jan Hutař, Ph.D a Mgr. Pavla Švástová), z Pravidel popisu (dostupná v sekci „Pravidla popisu pro digitalizaci“ na http://www.ndk.cz/standardy-digitalizace/metadata), Standardů pro souborové formáty a také pravidel pro identifikátory, která jsou částečně zachycena v sekci pro metadata a podrobněji v metodice pro užívání standardu URN:NBN – viz dále.

3 SIP – Submission Information Package. Tento i další pojmy jsou používány ve shodě s terminologickým slovníkem normy ČSN ISO 14721.

4 O významu prosazení jednotného standardu komplexněji *Návrh národní koncepce dlouhodobé ochrany digitálních dat pro knihovny* v kapitole 1.5 z roku 2014. Dostupný z http://ukr.knihovna.cz/koncepce-rozvoje-knihoven-cr-na-leta-2011-2015-/ (citováno 2017-09-30).

5 VISK 7. *VISK* [online]. [cit. 2017-10-01]. Dostupné z: http://visk.nkp.cz/visk-7.

Národní knihovny ČR a platformou pro výměnu poznatků a plánování dalšího vývoje Standardu je Formátový výbor NDK, který plní roli poradního orgánu.

Aktuálně umožňuje Standard NDK vytvářet SIP balíčky z digitalizovaných monografií a periodik včetně hudebnin a kartografických dokumentů, dále z digitálně vzniklých elektronických dokumentů a také z digitalizovaných audio dokumentů. Na následujících stranách se budeme věnovat jednotlivým částem Standardu NDK. Vedle toho, co bude zmíněno, je třeba upozornit, že součástí balíčků NDK (pro digitalizované monografie a periodika) jsou i soubory s rozpoznaným textem a soubory ALTO XML. I ty tvoří nedílnou součást struktury podle Standardu NDK.

Standardy pro metadata

Jak bylo řečeno, významnou součástí Standardu NDK je definice metadatové struktury.[6] Oblast knihovních dokumentů je v tomto směru poměrně specifická, protože v knihovnách existuje na rozdíl od dalších paměťových institucí značně komplexní systém popisu dokumentů založený na mezinárodních standardech. Jeho podstatu je třeba uchovat i v případě digitalizace, protože jinak hrozí ztráta informačního obsahu tak, jak byl zmíněn v úvodní pasáži. Nicméně metadatový popis neznamená samozřejmě jen věcné údaje. Standard NDK umožňuje v metadatové části široký popis všech nutných informací, které jsou nezbytné pro dlouhodobé uchovávání.

Primárně však nejde o jedno konkrétní schéma, které bychom mohli prohlásit za nejlepší. Z hlediska dlouhodobého uchovávání je především vysoce žádoucí, aby metadata byla zapsána v některém z rozšířených a dokumentovaných schémat. Těch existuje větší množství a záleží na rozhodnutí expertů, které doporučí pro konkrétní organizaci. Největší hrozbu představuje stav, kdy jsou metadata zapsána bez dokumentovaného schématu anebo odpovídají schématu, který si tvůrce dat sám vytvořil. Pokud za tvůrci nového schématu nestojí silná organizace, obvykle se stane, že udržování informace o předpisu selhává a ztrácí se jeho znalost, což pak v důsledku vede k neschopnosti správců dat pochopit informační obsah. Být schopen obsahově popsat každou metadatovou položku je nutné jednak z výše uvedeného důvodu, jednak kvůli organizaci dat v repozitáři. Obvykle bývá nutné převádět informace mezi jednotlivými schématy, zajistit mapování pro indexaci nebo při převádění do jiných systémů. Každá z těchto operací bezpodmínečně vyžaduje absolutní znalost jednotlivých metadatových elementů. Bez ní může docházet ke zmatkům, nepochopení nesené informace a ztrátě

6 Stadardy pro metadata. *NDK* [online]. [cit. 2017-10-01]. Dostupné z: http://www.ndk.cz/standardy-digitalizace/metadata.

smyslu dlouhodobého uchovávání a ochrany na logické úrovni. Bez znalosti metadatového schématu jsou metadata v balíčku pouze souborem informací, jejichž přesný význam nelze určit.

Principem Standardu NDK je samonosnost balíčků. K interpretaci informací v nich obsažených stačí jen Standard NDK. Žádné další kontextové informace ani data z repozitářového systému nebo jiných externích systémů nejsou třeba, vše v balíčku je popsané v daném schématu, jehož předpis musí být samozřejmě v repozitáři taktéž chráněn. V případě havárie systému repozitáře, produkčních systémů nebo knihovního katalogu by veškeré potřebné informace nutné pro plnou obnovu repozitáře a zachování informačního obsahu měly být obsaženy v AIP balíčcích. Samonosnost balíčku tvoří součást dlouhodobého uchovávání sama o sobě. I kdyby balíček vytvořený podle zásad Standardu NDK nebyl uložen v plnohodnotném repozitářovém systému, napomáhá jeho struktura uchování informačního obsahu právě samostatností na vnějších strukturách.

Z výše popsaných důvodů je tedy Standard NDK jediným povoleným schématem, podle něhož mohou být vyrobeny a následně uloženy SIP balíčky do LTP systému Národní knihovny ČR. Výhody to poskytuje všem zúčastněným stranám, producenti nebo firmy zaměřené na digitalizaci nemusí řešit sporné situace individuálně, mohou se spolehnout na metodickou podporu Národní knihovny ČR, jejich systémy mohou sloužit pro více zadavatelů a vzniká tím i poměrně silná uživatelská komunita umožňující řešení problémů ve vzájemné kooperaci. Zároveň vzniká i širší komunita, která rozumí logice a pravidlům Standardu a existuje tak do budoucna větší šance, že nedojde ke ztrátě uchovávané informace, což hrozí za situace, kdy pravidlům rozumí jen omezená skupina expertů. Sjednocením přístupu se také přirozeně snižují náklady na pořízení metadatového popisu. Pokud jde o druhou stranu procesu, tedy pracovníky repozitáře, tak homogenizace vstupujících balíčků jim usnadňuje práci v mnoha ohledech. K jejich úkolům patří porozumění vkládanému obsahu, což znamená, že u různorodých balíčků musí provádět analýzy kvůli poznání jejich struktury, případně k přípravě šablon, které mapují metadatový popis na systém indexu v repozitáři pro vyhledávání.

Vyhledávání podle co největšího počtu elementů patří k jedné z nejdůležitějších funkcí v rámci dlouhodobé správy, protože bývá potřeba identifikovat skupiny dat podle různých kritérií (hledání podle obsahových hesel, doby vzniku, platnosti autorských práv, konkrétních autorů, vydavatelů, dat vzniklých při výrobě – skener, software, pracovník apod.). Homogenizovaný popis, díky kterému kurátoři vědí, kde který údaj mohou hledat, představuje značnou výhodu snižující náklady na dlouhodobé uchovávání. Jde především o lidský čas, který mohou odborní pracovníci věnovat vlastní správě dat

a neplýtvat jim na snahu o porozumění heterogenní skupině dat vzniklých podle odlišných schémat. Bez užívání jednotného Standardu by bylo nutné mít v rámci NDK definované profily pro každý vstupující typ dat podle jednotlivých producentů, přičemž stejné typy by se u producentů mohly lišit (díky Standardu NDK stačí mít profil pro typ dat bez ohledu na producenta). V neposlední řadě je třeba zmínit, že jednotný Standard byl vytvořen i s ohledem na digitální knihovnu určenou pro zpřístupnění. Data vytvořená podle Standardu umožňují zobrazení v knihovně Kramerius bez dalších úprav.

Lze tedy konstatovat, že nastavení jednotného předpisu pro metadatovou strukturu pomocí Standardu NDK bylo poměrně striktním požadavkem, který se dotkl celé knihovnické komunity i spolupracujících firem. Na druhou stranu k jeho zavedení vedly logické důvody a i akceptace Standardu svědčí o potřebě podobného předpisu. Z hlediska dlouhodobého uchovávání představuje řešení s nejmenšími náklady a nejvyššími výhodami. Samozřejmě v sobě skrývá i jedno podstatné riziko, které by mohlo vzniknout při chybném nastavení Standardu. Chyba v něm obsažená by byla replikována na mnoha místech a následná oprava by byla nákladná. Nicméně široká komunita užívající Standard představuje dostatečnou prevenci nežádoucí chyby.

Přesuňme se nyní přímo k praktické podobě Standardu NDK v jeho metadatové části. Všechny využívané standardy vychází ze schémat udržovaných Library of Congress (dále LOC). Pro účely Standardu NDK jsou schémata LOC upravená, lokalizovaná pro české podmínky, ale jen v parametrech, které tyto standardy povolují (většinou to znamená částečné omezení u některých elementů, případně jejich zakázání nebo povolení jen jedné hodnoty). Základem celého popisu je kontejnerový formát METS, do něhož jsou zanořeny další schémata.[7] Podle typu dat jsou metadata vložena buď jen v jednom XML souboru, nebo případně ve více souborech, pak o dalších hovoříme jako o vedlejších METS souborech a hlavním METS souboru.

Popisná metadata jsou zapsána pomocí standardu MODS a Dublin Core.[8] Jejich obsah je de facto duplikovaný a toto dvojí užití je dané potřebou systémů, které s balíčky vytvořenými podle Standardu NDK pracují. Rozmanitější je oblast technických metadat, pro jejichž zápis je užit standard MIX v případě obrazových dat.[9] Pro zvukové dokumenty slouží standard AES57, který jako jediný využitý nevydává LOC, ale Au-

7 METS. *LOC* [online]. [cit. 2017-10-01]. Dostupné z: http://www.loc.gov/standards/mets/.

8 MODS http://www.loc.gov/standards/mods/ a Dublin Core http://dublincore.org/documents/dces/.

9 MIX. *LOC* [online]. [cit. 2017-10-01]. Dostupné z: http://www.loc.gov/standards/mix.

dio Engineering Society.[10] Další část technických a administrativních metadat je popsána ve standardu PREMIS, který je z hlediska dlouhodobého uchovávání zřejmě nejdůležitější.[11] Umožňuje podrobně popsat všechny aktivity, které byly s dokumentem spojené, kdo a za jakých okolností ho vytvořil, s jakými charakteristikami a jaké další operace nad ním byly vykonány. Celý komplex doplňuje schéma copyrightMD, které umožňuje zápis autorsko-právních metadat.[12]

Součástí metadatové části Standardu NDK je i popis povolené struktury SIP balíčku, který se skládá z několika adresářových struktur doplněných samostatnými soubory s metadaty. Tato struktura je podobně jako další části povinná a je nutné ji dodržovat. Jsou v ní definována pravidla pro umístění dat v rámci balíčku, jednotlivých metadatových položek, vazeb mezi hlavním a vedlejšími METS soubory, vazby na soubory s textovou vrstvou pomocí názvové konvence apod. Toto vše se následně promítá do strukturálních map obsažených vždy v hlavním METS souboru. Pomocí strukturálních map lze případně rekonstruovat vazby mezi jednotlivými soubory a pochopit celou strukturu balíčku jak po stránce prosté posloupnosti souborů, tak i z hlediska logických vazeb a významu. Pravidla pro podobu SIP balíčku je tak třeba vnímat se stejným významem jako vlastní metadatový předpis.

Souborové formáty

Vše, co bylo řečeno o významu standardizace v souvislosti s metadaty platí i pro souborové formáty. Základním předpokladem pro úspěšnost dlouhodobého uchování datových souborů je jejich vznik v rámci dokumentovaných formátů. Z hlediska činnosti NDK jde o základní požadavek. Všechny využité datové standardy musí být přesně a otevřeně popsány a soubory musí vzniknout podle nich. Odchylky nejsou možné. Ještě více než v metadatové části zde platí, že pracovníci repozitáře musí znát informace o vložených datech. Pro budoucí případné formátové migrace je nutné znát přesné specifikace dat, protože každá odchylka znamená v budoucnu nutnost nastavení speciálního procesu. Všechny odchylky od daného standardu znamenají potřebu oddělit taková data od ostatních a řešit je samostatně. I z tohoto důvodu je předepsán poměrně obsáhlý popis jednotlivých datových souborů v rámci technických metadat. V rámci standardizace NDK je nad nimi opakovaně prováděna validace spojená s identifikací souborových formátů.

10 http://www.aes.org/publications/standards/.

11 PREMIS. *LOC* [online]. [cit. 2017-10-01]. Dostupné z: http://www.loc.gov/standards/premis/.

12 CopyrightMD. *CDLIB* [online]. [cit. 2017-10-01]. Dostupné z: http://www.cdlib.org/groups/rmg/.

I v rámci určení doporučených respektive povolených souborových formátů platí to, co v předešlém případě. Standard NDK doporučil skupinu formátů vhodných pro dlouhodobé uchovávání, ale jde spíše o otázku k permanentní diskusi, než o jednou definitivně rozhodnuté doporučení. V rámci Standardu NDK jsou pro každý typ dat určené povolené formáty. Jen u nich lze aktuálně zaručit poskytování dlouhodobé ochrany na logické úrovni. Do budoucna se počítá s možností rozšíření na další formáty a to ve dvou směrech. Určitě je nutné předpokládat, že v budoucnu budou přijímány i další formáty, když se ukáže jejich potřeba nebo pokud se přijme rozhodnutí začít poskytovat logickou ochranu dalším typům. Druhou skupinu budou tvořit ty formáty, u kterých bude LTP úložiště Národní knihovny ČR poskytovat pouze ochranu na úrovni bitstream (půjde o data, kterým bude poskytnuta ochrana, ale právě jejich nestandardizovaná podoba neumožní poskytovat trvalou ochranu na logické úrovni). Aktuálně jsou však přijímána data jen v povolených formátech, respektive u některých typů jsou určeny povolené formáty a upravuje se vstup úložiště tak, aby je mohlo reálně začít přijímat.

Nejpočetněji zastoupeným typem dat v rámci úložiště jsou obrazové soubory, pro které je předepsán formát JPEG2000. Ten byl vybrán v počátku projektu NDK a zkušenosti s jeho využíváním potvrdily správnost jeho volby. Poskytuje vhodný poměr komprese a kvality se zachováním bezztrátovosti. Standard NDK neakceptuje jakýkoli JPEG2000, ale jen ten, který odpovídá přesnému profilu NDK.[13] Tento profil je průběžně testován a podrobován opakovaným přezkoušením. K jeho výrobě je doporučen nástroj Kakadu, případně OpenJPEG. Jak již bylo řečeno, přesný profil i využití doporučených nástrojů je podstatné z hlediska budoucích operací s daty. Každá odchylka může znamenat neúspěch formátové migrace, případně klade nároky na zjištění odchylky v rámci běžné správy dat. Pak musí následovat buď vrácení dat producentovi, anebo nalezení způsobu opravy dat v rámci repozitáře. Oboje je vysoce nežádoucí, zvyšuje finanční i lidské náklady dlouhodobého uchování a umožňuje vznik další nepředvídatelné chyby.

Ještě vyšší míra rizika je spojena s formáty digitálně vzniklých souborů. V případě dat, nad kterými bude vykonávat dlouhodobou ochranu úložiště Národní knihovny ČR, jde především o formáty elektronických knih. V tomto směru bylo rozhodnuto omezit preferované formáty na PDF/A ve verzi 1 a 2 a na EPUB ve verzi 2.0.1. V obou případech se jedná o kontejnerové formáty, které mohou obsahovat zanořené i další typy dat. Z tohoto důvodu byly z povolených formátů vyjmuty vyšší verze jmenovaných, protože mohou obsahovat např. audio soubory, animované obrázky a další entity, které není LTP úložiště aktuálně schopno dlouhodobě ochránit na logické úrovni. Důvodem této neschopnosti je právě rozmanitost možných typů dat a s tím spojená nemožnost důkladně popsat veškerý obsah

13 Standardy pro obrazová data. *NDK* [online]. [cit. 2017-10-01]. Dostupné z: http://www.ndk.cz/standardy-digitalizace/standardy-pro-obrazova-data.

daných souborů, který může být v různých proprietárních nebo nedostatečně dokumentovaných formátech. Důsledkem akceptování i takových souborů by byly následné neúspěšné formátové migrace a selhání celého procesu dlouhodobého uchovávání. Z hlediska dlouhodobého uchování představují kontejnerové formáty značné riziko, které je třeba mít na vědomí a alespoň částečně se mu vyvarovat administrativním opatřením a v budoucnu vývojem nástrojů, které budou schopny rozeznat i skryté formáty. Nestačí tedy mít zdokumentovaný kontejnerový formát, ale je nutné mít zdokumentované i všechny další užité formáty. V případě zvukových dokumentů je Standardem doporučen formát wav.[14]

Ani omezení souborových formátů v rámci Standardu NDK není samoúčelné. Úspěšné poskytování dlouhodobé ochrany digitálních dokumentů vyžaduje jejich důsledné pochopení a porozumění jejich schématům a principům jejich tvorby. Jde o vysoce specializované znalosti, které vyžadují detailně obeznámené odborníky. Týmy repozitářů jsou obvykle početně omezené a jejich odbornost může být dostatečně hluboká jen v určitých oblastech. Není v lidských ani finančních možnostech jakékoliv instituce poskytovat plnou a účinnou ochranu jakémukoliv typu digitálních dat. Správným řešením proto musí být omezení rozsahu činnosti na zvládnutelné oblasti. Ty přirozeně nemohou zůstat strnulé, ale musí se vyvíjet. Nicméně omezení na vstupu do repozitářů je nezbytné.

Identifikátory

Oblast identifikátorů je v rámci Standardu NDK stejně důležitá jako obě předešlé, nicméně nepodléhá tolika omezením. O významu identifikátorů pro dlouhodobou ochranu digitálních dokumentů již není pochyb. Dokazuje to jak praxe, tak teoretické studie zabývající se Digital Preservation.[15] Perzistentní identifikátory jsou v některých názorech tím hlavním svorníkem chráněných dat, které umožňují jejich dlouhodobou správu v různých systémech. Stále aktuálnější otázkou totiž je, jak přesně definovat digitální objekt, který může získat různorodou podobu. Robert Khan nabízí přístup, který chápe objekt jako v zásadě sadu bitů s unikátním identifikátorem, který by měl zajišťovat všechny operace

14 Standardy pro zvuková data. *NDK* [online]. [cit. 2017-10-01]. Dostupné z: http://www.ndk.cz/standardy-digitalizace/standardy-pro-zvukova-data.

15 Jako reprezentativní lze brát např. soubor příspěvků z konference iPres 2016, která otázku perzistentních identifikátorů významně tematizovala. https://ipr16.organizers-congress.org/frontend/organizers/media/iPRES2016/_PDF/IPR16.Proceedings_4_Web_Broschuere_Link.pdf (citováno 2017-09-28). Přehled některých evropských systémů identifikace digitálních dokumentů v článku CUBR, Ladislav, VAŠEK, Zdeněk et al. Srovnání vybraných národních identifikačních systémů užívajících identifikátory URN: NBN. *ProInflow: časopis pro informační vědy* [online], 2016, **8**(1) [cit. 2017-09-29]. ISSN 1804-2406. Dostupné z: http://www.phil.muni.cz/journals/index.php/proinflow/article/view/1220/1490.

s objektem. Součástí debat je i převažující názor, že vývoj se musí ubírat směrem k vývoji „chytrých" identifikátorů, které nebudou jen samy o sobě, ale budou schopny uživateli poskytnout i další informace o objektu, metadata, autorsko-právní informace.

Tyto nároky splňuje identifikátor URN:NBN, který je podle Standardu NDK spolu s identifikátorem UUID povinný.[16] Všechny další identifikátory, které odpovídají jakémukoliv standardu, jsou povolené.

Nástroje

Standard NDK přímo nedefinuje nástroje, které by měli tvůrci SIP balíčků používat. Z logiky věci vyplývá nutnost komunikace s Resolverem URN:NBN, který provozuje Národní knihovna ČR. Ten zajišťuje přidělování identifikátorů. Dále je doporučen nástroj Kakadu pro tvorbu obrazových souborů. Pro identifikaci souborových formátů a jejich validaci je možné využít software JHOVE a formátový registr PRONOM a na něm založené nástroje jako např. DROID (případně FIDO).

Validace vytvořených dat podle Standardu NDK je jedním z důležitých prvků spojených s užíváním a uplatňováním Standardu. Z toho důvodu vznikl Komplexní validátor NDK, který umožňuje validace jak metadatové struktury, tak vlastních dat.[17] Jde o lokálně instalovatelný software, který slouží jako metodická pomůcka pro uživatele Standardu NDK. S jeho pomocí mohou validovat svá data, v případě nalezení chyby jim poskytne zpětnou vazbu, kde se vyskytla chyba a jakou doporučuje nápravu. V konečné fázi je o chybách samozřejmě nutné diskutovat s pracovníky Národní knihovny ČR. Zatím je k dispozici nastavení pro digitalizované dokumenty, do budoucna se počítá s využitím i pro další výše popsané typy dat. Komplexní validátor je tak třeba chápat jako jistou součást Standardu NDK. Umožňuje vytvářet data podle Standardu všem jeho uživatelům.

Pokud jde o nástroje pro tvorbu SIP balíčků podle Standardu NDK, lze konstatovat, že v ČR existují komerční i otevřené nástroje. Svými softwary, které dokáží splnit nároky Standardu, disponují komerční společnosti, které poskytují digitalizační služby kni-

16 Užívání identifikátoru se řídí pravidly certifikované metodiky VAŠEK, Zdeněk a kol. *Metodika pro přidělování a správu životního cyklu unikátních perzistentních identifikátorů digitálních dokumentů podle standardu URN: NBN* [online]. 2015 [cit. 2017-10-01]. Dostupné z: http://aleph.nkp.cz/F?func=service&doc_library=KKL01&local_base=KKL&doc_number=000083859&line_number=0001&func_code=WEB-FULL&service_type=MEDIA.

17 Komplexní validátor NDK. *GitHub* [online]. [cit. 2017-10-01]. Dostupné z: https://github.com/NLCR/komplexni-validator/wiki.

hovnám. V tomto případě plní Standard roli centrálního předpisu, díky kterému je přesně definovaný výstup poskytované služby, čímž odpadají diskuse o její kvalitě. Pomocí standardizace výstupu je také umožněn větší výběr knihoven mezi ověřenými dodavateli. Vedle komerčních produktů je k dispozici též nástroj ProArc, který je volně dostupný a lze ho využít pro tvorbu SIP balíčků podle Standardu.[18] Národní knihovna ČR využívá vlastní komerčně dodaný nástroj, který však byl upraven podle jejích potřeb.

Závěr

Standard NDK se od počátku svého užívání prosadil v praxi knihoven ČR. Od roku 2012 nabízel doporučení pro vytváření SIP balíčků pro digitalizáty monografických publikací a periodik. Aktuálně je připravován na rozšíření o další typy digitálních dokumentů včetně ebornů. Jednotný standard pro produkci, který umožňuje následné poskytování dlouhodobého uchovávání digitálních dokumentů, lze považovat za výrazný úspěch celé odborné komunity. Přináší úspory při produkci i díky synergickým efektům při jeho využívání. Jeho struktura patří ke komplexnějším řešením, na druhou stranou právě díky své struktuře mají SIP balíčky vlastnosti, které napomáhají dlouhodobému uchování. Klíčová v celé standardizaci je role Národní knihovny ČR, která má za standard odpovědnost a odpovídá za jeho udržování, čímž poskytuje službu dalším knihovnám. Dodržování pravidel Standardu NDK je základní podmínkou pro zvládnutí otázky ochrany digitálních dokumentů, odchylky od Standardu by ji prodražily nebo rovnou znemožnily.

Odkazy a literatura

BORGHOFF, Uwe M. et al. *Long-term preservation of digital documents: principles and practices*. Berlin: Springer, 2005. xv, 274 s. ISBN 3-540-33639-7.

CUBR, Ladislav. *Dlouhodobá ochrana digitálních dokumentů*. 1. vyd. Praha: Národní knihovna České republiky, 2010. 154 s. ISBN 978-80-7050-588-5.

CUBR, Ladislav, VAŠEK, Zdeněk et al. Srovnání vybraných národních identifikačních systémů užívajících identifikátory URN: NBN. *ProInflow: časopis pro informační vědy* [online], 2016, **8**(1) [cit. 2017-09-29]. ISSN 1804-2406. Dostupné z: http://www.phil.muni.cz/journals/index.php/proinflow/article/view/1220/1490

18 ProArc. *GitHub* [online]. [cit. 2017-10-01]. Dostupné z: https://github.com/proarc/proarc/wiki.

GIARETTA, David. *Advanced digital preservation*. Berlin: Springer, 2011. xxii, 510 s. ISBN 978-3-642-16808-6.

HUTAŘ, Jan, 2012. *Digitalizace, popis pomocí metadat a jejich formáty*. Praha, 244 s. Disertační práce. Univerzita Karlova v Praze, Ústav informačních studií a knihovnictví.

HUTAŘ, Jan a MELICHAR, Marek. Dlouhodobá archivace digitálních dat – od teoretických úvah k praktické realizaci?. *Knihovna: knihovnická revue*, 2015, **26**(2), s. 58-68. ISSN 1801-3252. Dostupné také z: http://knihovnarevue.nkp.cz/aktualni-cislo/knihovny-a-informace/dlouhodoba-archivace-digitalnich-dat-2013-od-teoretickych-uvah-k-prakticke-realizaci

Návrh národní koncepce dlouhodobé ochrany digitálních dat pro knihovny [online]. UKR, 2014 [cit. 2017-10-01]. Dostupné z: http://ukr.knihovna.cz/koncepce-rozvoje-knihoven-cr-na-leta-2011-2015-/

Standard NDK. *NDK* [online]. [cit. 2017-10-01]. Dostupné z: http://www.ndk.cz/standardy-digitalizace/standardy-digitalizace-1

Standards LOC. *LOC* [online]. [cit. 2017-10-01]. Dostupné z: https://www.loc.gov/librarians/standards

Systémy pro přenos dat a informací z kosmického prostoru – Otevřený archivační informační systém – Referenční model = Space data and information transfer systems – Open archival information system (OAIS) – Reference model = Systèmes de transfert des informations et données spatiales – Système ouvert d'archivage d'information (SOAI) – Modèle de référence: ČSN ISO 14721: schváleno v červnu 2012 ve Washingtonu, DC, USA. Druhé vydání. Praha: Úřad pro technickou normalizaci, metrologii a státní zkušebnictví, 2014. 97 stran.

VAŠEK, Zdeněk a kol. *Metodika pro přidělování a správu životního cyklu unikátních perzistentních identifikátorů digitálních dokumentů podle standardu URN: NBN* [online]. 2015 [cit. 2017-10-01]. Dostupné z: http://aleph.nkp.cz/F?func=service&doc_library=KKL01&local_base=KKL&doc_number=000083859&line_number=0001&func_code=WEB-FULL&service_type=MEDIA

ZENG, Marcia a QIN, Jian. Metadata. 2nd edition, 2016. 822 s. ISBN 978-1-55570-965-5

SIP balík – Ako na to?

Ako má vyzerať štruktúra SIP balíka pre CDA

Jaroslav Kamenský, Ľubomír Hribík, Tempest, a. s.

Abstrakt
Predložená práca sa snaží svojim obsahom priblížiť problematiku vytvárania SIP balíkov. Autori sa snažia poukázať na fakt, že pozornosť pri vytváraní SIP balíka treba venovať nielen technickým prostriedkom (HW resp. „balíčkovacia“ aplikácia) ale najmä správnej tvorbe metadát a ich praktickej využiteľnosti v budúcnosti.

Úvod

Centrálny dátový archív (CDA) má uzatvorené zmluvy – dohody o zverení obsahu na dlhodobú archiváciu s viacerými Pamäťovými a fondovými inštitúciami (PFI). Od týchto PFI prijíma dáta z ich vlastných digitalizačných projektov vo forme tzv. SIP balíkov. Štruktúra týchto balíkov ako aj spôsob akým je popísaný ich obsah sú pevne dané a každá PFI musí tieto pravidlá tvorby SIP balíkov dodržať inak ich dáta nebudú uložené v CDA.

Niektoré PFI používajú v procese digitalizácie softvér na báze workflow. Ide automatizovaný prípadne polo-automatizovaný proces pozostávajúci zo sekvencie krokov, na ktorého výstupe vzniknú tzv. digitálne objekty (DO). Tieto digitálne objekty sú väčšinou vo forme adresárovej štruktúry so súbormi, ktoré digitalizačná linka resp. softvér vyprodukoval. Metadáta o objektoch si PFI evidujú v špecializovaných informačných systémoch, z ktorých dáta musia buď exportovať, alebo ručne prepisovať do textových alebo iných popisných súborov. Následne takto zozbierané dáta musia použiť na vytvorenie SIP balíka v štruktúre akú predpisuje CDA. Niektoré PFI majú také informačné systémy, alebo dodaný workflow softvér pre digitalizáciu, ktoré dokážu na svojom výstupe vygenerovať dáta vo forme tzv. PSP balíka. V tomto prípade sa použije softvér, ktorý dokáže konvertovať PSP balíky na SIP balíky pre CDA.

Príprava

Technické prostriedky na generovanie SIP balíkov nie sú jediné, čo je nutné zabezpečiť na bezproblémový a efektívny proces ukladania dát do CDA. Inštitúcia vystupujúca v roli vkladateľ do CDA by mala mať prostriedky a prípadne aj nastavené interné procesy aby vedela pokryť nasledovné oblasti, ktoré sa v praxi ukázali ako kľúčové pre zabezpečenie efektívneho vkladania do CDA:

1. Vytvoriť (diskový) priestor pre dočasné úložisko SIP balíkov
   Pre efektívnejšiu prácu je lepšie oddeliť výstupy z digitalizácie od miesta, kde budú uložené SIP balíky, pripravené na odoslanie do CDA a to preto, že SIP balík má predpísanú štruktúru a môže obsahovať výhradne len súbory vo vopred dohodnutých formátoch.

2. Pripraviť sa na veľké objemy a/alebo veľké počty balíkov
   Je dobré predpokladať, že vklady do CDA nebudú prebiehať po jednotlivých balíkoch, ale bude to skôr vo forme veľkej dávky viacerých balíkov. CDA zaviedla pre tieto účely pojem kampaň, ktorý reprezentuje proces vkladu množiny SIP balíkov od jedného alebo viacerých vkladateľov. Kampaň je pridelená operátorovi v CDA, ktorý je za ňu zodpovedný a riadi celý proces jej spracovania.

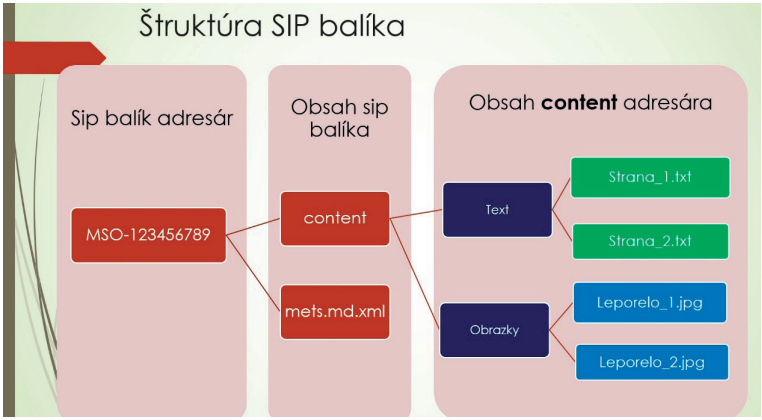
3. Rozšíriť evidenciu o CDA identifikátory
   Každý balík musí mať jedinečný identifikátor v systéme CDA. Balíky, ktoré dodávajú vkladatelia sú SIP balíky a balíky, ktoré sú uložené v CDA archíve sú AIP balíky. Systém CDA pozná väzbu medzi SIP_ID a AIP_ID a umožňuje vyhľadať balíky podľa oboch identifikátorov. Dôvod prečo by si mal vkladateľ viesť rovnakú evidenciu je ten, ak by potreboval na tieto identifikátory naviazať svoje interné procesy. Ide napríklad o procesy, alebo evidenciu chybných balíkov, prípadne evidenciu stavu prípravy balíkov pre CDA a ich úspešnosť/neúspešnosť pri vklade do CDA.

4. Implementovať proces náhrady SIP balíka
   Každý SIP balík má jedinečný identifikátor, ktorý si systém CDA zaeviduje po prijatí balíka. Oprava SIP balíka, ktorý skončil v chybe pri spracovaní z dôvodu nepovoleného obsahu, alebo chybne uvedených metadát o balíku, nie je v CDA povolená. Preto je vhodné implementovať proces náhrady SIP balíka a rozšíriť internú evidenciu o SIP_ID, aby bolo možné sledovať koľko balíkov bolo vygenerovaných a odoslaných do CDA z jedného zdroja dát a ako ich spracovanie

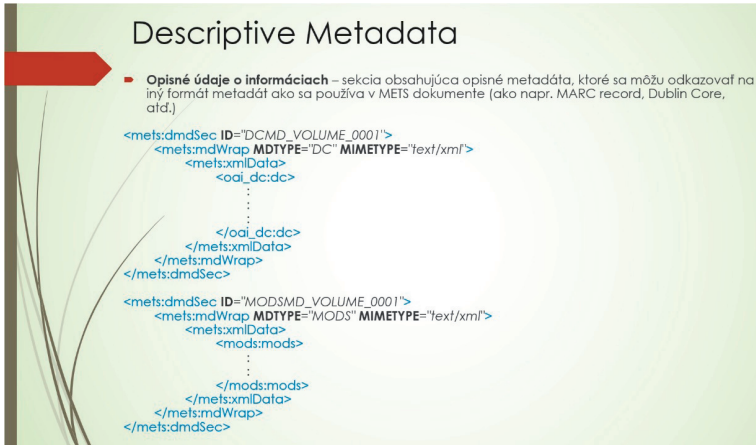
skončilo. Z úspešne spracovaného SIP balíka vznikne AIP balík, ktorému systém CDA pridelí AIP_ID.

5. Implementovať procesy pre prácu so súborovými formátmi
   Najčastejšou chybou v procese spracovania SIP balíkov v CDA je rozdielnosť formátu súboru oproti zoznamu formátov uvedených v dohode medzi CDA a vkladateľom. Vo väčšine prípadov ide o neúmyselnú chybu spôsobenú zariadením, alebo softvérom ktorý dáta pri digitalizácii produkuje. Z tohto dôvodu je dobré implementovať procesy pre kontrolu správnosti formátov, alebo zoznam formátov uvedený dohode s CDA mať záväzný aj pre proces digitalizácie. Dôležité je si uvedomiť, že v dohode sa neuvádza iba názov formátu súboru (napr. PDF, DOC, TXT, JPG, WAV) ale aj jeho verzia resp. verzie. Chyba pri spracovaní potom nastáva, ak sa v SIP balíku vyskytujú dáta v inej verzii súborového formátu, ako bol dohodnutý. Ak chce vkladateľ predísť týmto chybám, tak najspoľahlivejší spôsob je implementovať formátovú identifikáciu pre kontrolu SIP balíkov.

Štruktúra SIP balíka a štruktúra popisných metadát



Obr. 1 – Štruktúra SIP balíka



Obr. 2 – Štruktúra popisných metadát v štandarde METS

Praktická ukážka priamo na konferencii.

Tvorba SIP balíkov SW produktom CDA UKB

Roman Král, Univerzitná knižnica v Bratislave

Abstrakt
Stručná charakteristika programu EXCELMETS, ktorý slúži na tvorbu SIP balíkov, pre potreby PFI spolupracujúcich s CDA. Program je postavený na platforme Microsoft Excel. Program obsahuje aj niektoré ďalšie funkcionality, ako napr. evidenciu, testovanie alebo hromadné spracovanie súborov.

Program, ktorého interný názov je „EXCELMETS“ je program na vytváranie SIP balíkov (SIP-submission information package). História jeho vývoja bola taká, že Univerzitná knižnica (UKB) potrebovala vkladať balíky do Centrálneho Dátového archívu (CDA) a hľadala spôsob, ako na to. Mali už vytvorené zoskenované data z kníh a časopisov na diskových poliach. Dostupný balíčkovač MES vytvorený konzorciom TMG vykazoval v tom čase chybu, ktorej odstraňovanie sa z rôznych príčin časovo naťahovalo. Nebol teda pre UKB v tom čase použiteľný. Na vytvorenie informačného balíka pre vklad (SIP=submission information package) vhodného pre CDA bolo treba dodržať určité špecifikácie. V prvom rade bolo pre každý balík vytvoriť popisný súbor podľa štandardu METS (metadata encoding and transmission standard) – v CDA je to takzvaný mets_ md.xml súbor. Adresáre a súbory bolo treba usporiadať do správnej štruktúry elektronicky podpísať popisný dokument a nakoniec celý balík zabaliť do jedného súboru.

Ako najvhodnejšie sa nám javilo vytvoriť v CDA vlastný balíčkovač a to v aplikácii Microsoft Excel. Microsoft Excel poskytuje grafické rozhranie pre prácu s tabuľovými údajmi a má integrované progamátorské prostredie – Visual Basic. Tak vznikla myšlienka vytvárania SIP balíkov v exceli. Tento program dostal meno Excelmets ako spojenia platformy a toho najdôležitejšieho v balíku – popisného mets suboru.

Aplikácia EXCELMETS obsahuje 4 zošity (sheets) – DATA, CONF,TEST a BATCH. Tieto sú v pozadí navzájom poprepájané. Jednotlivé akcie sa v aplikácii púšťajú pomocou tlačítiek v danom zošite.

Hlavný je zošit DATA v ktorom sa zadáva identifikačné číslo vkladového balíka – SIPid ako aj sa v tomto zošite zozbierávajú popisné údaje vo formáte Dublin core. Tiež sa tu zadáva umiestnenie adresára, z ktorého sa má vytvoriť SIP balík. Potom už je na užívateľovi, či sa rozhodne balík hneď vytvoriť, alebo zozbierané údaje presunie do zošitu BATCH. Samotné vytváranie balíkov je totiž časovo náročné. Pre každý súbor sa vypočítava kontrolná suma (md5) a nakoniec sa balík skompresuje do ZIP formátu. Užívateľ môže použiť batchové spracovanie, keď napríklad cez deň v práci zozbierava údaje k balíkom a na noc pustí samotné generovanie.

V zošite CONF sa zadávaju konfiguračné údaje. Je tam napriíklad tabuľka prevodu súborového typu na mime-type. Zošit CONF je nezávislý a umožňuje rýchly test súborov programmi droid a jhove. Práve tieto programy používa CDA pri vklade balíka pre identifikáciu a validáciu súborov. A posledný zošit -BATCH umožňuje spustiť hromadné spracovanie predpripravených balíkov.

Počas behu spúšťa EXCELMETS rôzne aplikácie tretích strán. V prípade potreby sa púšťa java, openssl, 7z, iconv, droid, jhove a iné.

Logovanie celého procesu vytvárania balíku sa uskutočňuje zápisom do textového log súboru, ako aj do externej excelovskej tabuľky, nazvanej EVIDENCIA. Aplikácia EVIDENCIA umožňuje ku každému vloženému SIP balíku získať stav spracovania v CDA a tiež aj názov výsledného AIPu (Archival Information Package).

Výkonnosť celej aplikácie je dostatočná a na moderných PC dosahuje až 1TB dát v balíkoch za deň.

Zoznam autorov

Milan Rakús
Univerzitná knižnica v Bratislave

Zuzana Kvašová
Národní knihovna České republiky

Monika Péková
Ministerstvo vnútra SR

Darek Paradowski
National Library of Poland

Miklós Lendvay
National Széchényi Library, Hungary

Martin Lhoták
Knihovna Akademie věd ČR

Peter Selecký
Národné osvetové centrum

Ladislav Cubr
Národní knihovna České republiky

Szabolcs Dancs
National Széchényi Library, Hungary

Juraj Strnisko
Univerzitná knižnica v Bratislave

Zdeněk Vašek
Národní knihovna České republiky

Jaroslav Kamenský, Ľubomír Hribík
Tempest, a. s.

Roman Král
Univerzitná knižnica v Bratislave

CDA 2017

Výmena skúseností z prevádzky a budovania LTP archívov

Vydala Univerzitná knižnica v Bratislave
Prvé vydanie. Počet strán 132.
Sadzba: DOLIS, s.r.o., Bratislava
Tlač: DOLIS, s.r.o., Bratislava

ISBN 978-80-89303-58-8
ISSN 2453-9406

ISBN 978-80-89303-58-8
ISSN 2453-9406