UNIVERZITNÁ KNIŽNICA V BRATISLAVE

CDA 2016

Formátové výzvy LTP

Zborník príspevkov z 1. medzinárodnej konferencie o dlhodobej archivácii

univerzitná knižnica
v bratislave

Bratislava, 2016

UNIVERZITNÁ KNIŽNICA V BRATISLAVE

CDA 2016

Formátové výzvy LTP

Zborník príspevkov z 1. medzinárodnej konferencie
o dlhodobej archivácii

univerzitná knižnica
v bratislave

Bratislava, 2016

© Univerzitná knižnica v Bratislave, 2016

Zostavovateľka
Mgr. Lucia Kelemenová

Autori príspevkov
Ing. Milan Rakús
Mgr. Bibiána Žigová
Mgr. Jan Hutař, Ph.D.
PhDr. Ladislav Cubr
Bakk. techn. Peter Bubestinger
Bc. Andrej Bizík
Mgr. Jaroslav Kvasnica

Obálka a grafický návrh
DOLIS, s.r.o., Bratislava

CIP SR

CDA 2016 [online] : Formátové výzvy LTP : zborník príspevkov z 1. medzinárodnej konferencie o dlhodobej archivácii: Bratislava, 10. 11. 2016 / zost. Lucia Kelemenová ; obálka a graf. návrh DOLIS, s.r.o. – 1. vyd. – Bratislava : Univerzitná knižnica v Bratislave, 2016

LTP archívy. Centrálny dátový archív. Dlhodobé dôveryhodné digitálne úložisko. Formátová stratégia. Formáty súborov. Kontajnerové formáty súborov. Kodeky.

ISBN 978-80-89303-53-3
ISSN 2453-9406

Obsah

ALOJZ ANDROVIČ – SILVIA STASSELOVÁ
Úvod 5

MILAN RAKÚS
CDA a Formátová stratégia CDA 7

BIBIÁNA ŽIGOVÁ
Formáty a LTP v Európe 20

JAN HUTAŘ
Identifikace formátu – jednorázový nebo opakovaný proces? 35

LADISLAV CUBR
Formátová strategie LTP úložiště NK ČR 44

PETER BUBESTINGER
File formats for audiovisual preservation: How to choose? 58

ANDREJ BIZÍK
Webový archívny formát WARC 81

JAROSLAV KVASNICA
WARC 1.1 je skoro tady – co přinese nová verze? 93

Úvod

Trvalé a spoľahlivé uchovávanie a sprístupňovanie informácií, poznatkov a skutočností v najrôznejších formách, ktoré je základným a prirodzeným poslaním pamäťových inštitúcií, nadobudlo v ostatných rokoch nový virtuálny a globálny rozmer. Digitalizácia obsahu, elektronizácia komunikácie a kybernetizácia spracovania údajov predstavujú nové výzvy najmä pre depozitné knižnice a archívy, ako aj mnohých ďalších aktérov v uvedenej oblasti. Dôveryhodná dlhodobá digitálna archivácia je štandardizovaný súbor organizačných, procesných a technologických predpokladov a opatrení na zabezpečenie dlhodobej ochrany digitálnych objektov a dátových štruktúr.

Univerzitná knižnica v Bratislave prijala v závere roka 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 vyústila už koncom roku 2012 do vkladu prvého archívneho balíka. Dnes, na sklonku roku 2016, má za sebou CDA dva roky reálnej, náročnej, ale aj úspešnej prevádzky.

Informačné systémy pre dlhodobú archiváciu sa dnes opierajú najmä o konsolidovaný súbor noriem, odporúčaní a príkladov dobrej praxe. Rutinná prax však spravidla na dennej báze prináša so sebou aj nové problémy, ktoré sú zároveň výzvou pre nové skúsenosti. Fenomén digitálnych údajov, objektov a štruktúr nie je uzavretou záležitosťou, vyvíja a mení sa v čase vďaka výskumnej a vývojovej tvorivosti, súbežne s rozvojom technológií a nárastom požiadaviek praxe. Je aktuálnym predmetom experimentov, výskumu a vývoja, na ktorý treba reagovať aj v komplexných súvislostiach a špecifických požiadavkách dlhodobej dôveryhodnej archivácie. Osobitnou a mimoriadne zaujímavou témou v uvedenej oblasti sú digitálne formáty, ktoré vo svojej dnes už „historickej“ rozmanitosti a súčasnej dynamike vyžadujú priebežnú pozornosť a odpovedajúcu odozvu v prevádzke, a najmä pri rozvoji dôveryhodných LTP systémov. Formátová politika, resp. stratégia je neodmysliteľnou súčasťou prevádzkovej dokumentácie dôveryhodného dlhodobého archívu. Jej vytvorenie predpokladá na jednej strane hlbokú znalosť vnútorných atribútov digitálnych entít, a na druhej strane dostatok praktických skúseností, vrátane komplexného pohľadu na doterajší a anticipovaný ďalší vývoj v danej oblasti. Cieľom organizátorov konferencie CDA 2016: Formátové výzvy LTP je prispieť k úrovni poznania v danej oblasti formou vybraných nosných príspevkov, a tiež prostredníctvom diskusií a výmeny praktických a teoretic-

kých poznatkov a názorov v medzinárodnom kontexte. Pozornosť, ktorú touto formou venujeme témam dlhodobej ochrany „digitálnych“ znalostí je zároveň príspevkom k napĺňaniu ambície Univerzitnej knižnice v Bratislave, najstaršej vedeckej knižnice na Slovensku, v oblasti vedeckej a výskumnej činnosti. Veríme, že aj vďaka konferencii na pôde Univerzitnej knižnice v Bratislave vznikne iniciatíva budúcej systematickej a dlhodobej spolupráce zainteresovaných expertov a inštitúcií.

1. medzinárodná konferencia CDA 2016: Formátové výzvy LTP sa uskutočnila dňa 10. 11. 2016 v Univerzitnej knižnici v Bratislave. Príspevky v zborníku sú zoradené tak, ako odzneli na konferencii.

V mene organizátorov konferencie

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

Centrálny dátový archív a formátová stratégia 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 dva roky prevádzky (2015 – 2016). Dva roky je dostatočná doba na to, aby sa doladili a overili základné procesy vkladu a výberu dát do a z archívu. Dva roky je dostatočná doba na to, aby boli identifikované problémy, ktoré so sebou prináša dlhodobá ochrana kultúrneho dedičstva v digitálnej forme v prostredí dôveryhodného úložiska. Príspevok sa pokúša, po stručnom úvode, identifikovať problémy Centrálneho dátového archívu v oblasti formátovej ochrany dát a v niektorých prípadoch načrtáva možnosti ich riešenia.

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 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 je to lokalita CDA-A a v Martine lokalita CDA-B. Obe lokality fungujú autonómne a každá z nich dokáže plnohodnotne zastúpiť funkciu druhej v prípade poruchy alebo odstávky. Okrem dvoch aktívnych lokalít disponuje CDA aj pasívnym skladom archivačných médií v lokalite CDA-C, ktorá sa nachádza v Bratislave. 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) (https://www.crz.gov.sk/index.php?ID=2288584&l=sk), Čiastková zmluva na poskytovanie služieb podpory NON IKT CDA (https://www.crz.gov.sk/index.php?ID=2586212&l=sk), 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 dva roky prevádzky (2015 – 2016) [4].

V rámci programu OPIS PO2 sa súbežne s realizáciou projektu CDA realizovalo aj 5 nosných digitalizačných projektov (Digitálna knižnica a digitálny archív, Digitálna galéria, Digitálne múzeum, Digitálny pamiatkový fond a Digitálna audiovízia) a niekoľko dopytových projektov [2]. Riešitelia týchto projektov predstavujú určené spoločenstvo v súvislosti s CDA.

2. Centrálny dátový archív – HW riešenie, SW – riešenie, základné procesy, organizačné a personálne zabezpečenie

HW riešenie CDA

HW riešenie CDA možno rozdeliť do týchto šiestich relatívne samostatných celkov: LAN infraštruktúra, SAN infraštruktúra, Serverová farma, Diskové pole, Komunikačná magneticko-pásková knižnica a Archívna magneticko-pásková knižnica.

LAN infraštruktúra. Tvoria ju sieťové prvky, ktoré zabezpečujú pripojenie CDA do Internetu a do internej LAN siete. Využíva prvky: Core prepínače HP 10504 (2 ks per lokalita), Firewally HP JG213A (2 ks per lokalita) a Routery HP JG311A (2 ks per lokalita).

SAN infraštruktúra. Tvoria ju optické prepínače, ktoré spájajú servery, diskové pole a páskové knižnice. CDA využíva prepínače SAN 80B-4 (2 ks per lokalita) a SAN 24B (2 ks per lokalita).

Serverová farma. Na spracovanie dát a výpočtov slúžia hlavné servery IBM Power 770 (2 ks per lokalita), typové označenie x 64. Každý server má 48 procesorov, 48 jadier procesorov, frekvenciu 3,3 GHz, pamäť RAM o veľkosti 384 GB (DDR3). Servery HP, typové označenie x 86, a x 64, (11 per lokalita) z toho 8 ks HP DL360p G8, 2 ks HP DL580 G7, 1 ks HP DL320G6 využíva CDA na podporu zabezpečenia prevádzky aplikácií.

Diskové pole. Slúži ako dočasné úložisko pre systém CDA. CDA využíva celkovo 12 diskových kontrollerov, 20 expanzných jednotiek), 608 ks HDD (600 GB), 42 ks SSD (400GB), 80 ks HDD (900 GB), v každej lokalite. Celková kapacita diskového poľa je aktuálne cca 220 TiB (per lokalita), spolu obe lokality 440 TiB.

Komunikačná magneticko-pásková knižnica. Slúži na vstupno-výstupné operácie s dátami, ktoré prichádzajú od určeného spoločenstva, alebo sú pre toto spoločenstvo určené. Knižnica IBM System Storage TS3500 Tape Library má nainštalovaných 10 ks LTO5 drivov (typ TS1050) a 4 ks LTO6 drivov (typ TS1060). V každej lokalite je umiestnená jedna takáto knižnica.

Archívna magneticko-pásková knižnica. Slúži ako dlhodobé úložisko dát. Knižnica IBM System Storage TS3500 Tape Library má nainštalovaných 12 ks drivov E07 (typ TS1140) a 1 drive E08 (typ TS1150). Kapacita knižnice je 25 PB. V každej lokalite je umiestnená jedna takáto knižnica.

SW riešenie CDA

Systém je principiálne koncipovaný ako redundantná Linux/Unix serverová farma, ktorá pozostáva z týchto súčastí: Operačné systémy, Komponenty IBM, Tempest – Databázový Server, Tempest – Aplikačný Server, Tempest – System Directory, Java impex a Framework.

Operačné systémy. AIX (produkcia), Redhat (produkcia), Windows (podporné služby)

Komponenty IBM. LTFS (Linear Tape File System), GPFS (General Parallel File System), HSM (Hierarchical Storage Management), TSM (Tivoli Storage Management)

Tempest – Databázový Server (Tempest-DS). Produkt, ktorý poskytuje spojenie troch rozdielnych prístupov realizácie databázových systémov. Je zložený z riešenia pre relačné spracovanie dát, nerelačné spracovanie dát v štruktúre kľúč/hodnota a špecializovaného databázového systému zameraného na lexikografické spracovanie dát. Tempest-DS je realizovaný ako komplexná kastomizovaná databázová aplikácia s využitím funkcionality Oracle MySQL Enterprise Edition, Apache HBase a Apache SOLR.

Tempest – Aplikačný Server (Tempest-AS). Platforma pre prevádzku výkonných komponentov CDA – komponentov, ktoré implementujú riadenie procesov CDA (ingescia, diseminácia, dlhodobé uchovávanie). Tempest-AS je realizovaný ako komplexná kastomizovaná aplikácia na platforme Java s využitím funkcionality Apache Tomcat, Apache Httpd, mod_proxy, mod_jk, h2db a mysql-connector.

Tempest – System Directory (Tempest-SD). Kombinuje implementáciu LDAPv3 (Lightwight Directory Access Protocol), DSMLv2 (Directory Service Markup Language), ako nástrojov na poskytovanie adresárových služieb a implementáciu centrálnej autorizačnej služby pre účely SSO (Single-Sign-On).

Java impex. Komponent, ktorý zabezpečuje import/export operácie medzi servermi a magneticko-páskovými knižnicami.

Framework. Aplikačný komponent navrhnutý na implementáciu biznis procesov. Základným prvkom CDA frameworku je úloha. Príkladom je úloha vkladu konkrétneho SIP balíčka do archívu CDA, pričom úloha je realizovaná ako postupnosť niekoľkých krokov.

Základné procesy CDA

CDA má zvládnuté procesy vkladu, synchronizačného vkladu a výberu informačných balíkov z CDA. Procesy vkladu a výberu informačných balíkov z CDA sa môžu realizovať online alebo s využitím magnetických pások typu LTO.

CDA naštartoval permanentné procesy dlhodobého uchovávania zvereného obsahu, bitovú ochranu (kontrolu integrity), formátovú ochranu a synchronizáciu lokalít.

Organizačná štruktúra a personálne zabezpečenie CDA

CDA je organizačne začlenený do organizačnej štruktúry UKB do Úseku elektronizácie a integrácie ako Odbor Centrálneho dátového archívu (OCDA). Tvoria ho tri oddelenia: Oddelenie informačných procesov, Oddelenie informačných systémov a Oddelenie informačných technológií. Prevádzku a rozvoj CDA zabezpečuje 12 zamestnancov UKB.

3. Centrálny dátový archív ako LTP archív

Medzi dátovými archívmi majú osobitné postavenie dôveryhodné dlhodobé úložiská, často- krát označované ako LTP (Long Term Preservation) archívy. Predpokladá sa, že informácie v nich budú uložené veľmi dlho, mali by byť stále čitateľné a mali by byť neustále prístupné používateľovi. Úloha nie je jednoduchá, pretože 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 zabezpečené a pod.

Existuje niekoľko metód auditu a certifikácie LTP archívov. Niektoré z nich, používané v európskom priestore [5], sú uvedené nižšie:

DSA (Data Seal of Approval) Pečať kvality digitálneho repozitára
DIN 31644 Information and documentation – Criteria for trustworthy digital archives
STN ISO 16363:2014 : Systémy prenosu vesmírnych údajov a informácií. Audit a certifikácia dôveryhodných digitálnych úložísk
DRAMBORA (Digital Repository Audit Method Based on Risk Assessment) (http://www.repositoryaudit.eu/)

Centrálny dátový archív bol projektovaný ako LTP archív. CDA je finančne zabezpečený (udržateľný) minimálne na obdobie udržateľnosti, má silné HW a SW riešenie, je organizačne a personálne pokrytý, má zvládnuté procesy tvorby a uchovávania archívnych informačných balíkov, má presne definované určené spoločenstvo používateľov, má certifikát kvality podľa normy STN ISO/IEC 27001:2013 [6], ktorým sú pokryté požiadavky na riadenie bezpečnostných rizík, má vypracovanú Politiku uchovávania údajov a spustil proces permanentného auditu a certifikácie CDA podľa normy STN ISO 16363:2014 [7]. CDA je navrhnutý tak, aby v maximálnej možnej miere využíval otvorené štandardy. CDA je implementovaný v súlade s ISO štandardom STN ISO 14721:2014 (OAIS) [3].

4. Formátová rozmanitosť

Formát súboru (súborový formát, typ súboru) je konkrétny spôsob kódovania informácií s cieľom ich uchovania v počítačovom súbore. Existujú rôzne druhy formátov pre rôzne druhy informácií. V rámci každého formátového typu existuje spravidla niekoľko rozdielnych, častokrát konkurenčných formátov. Formát súboru je daný jeho špecifikáciou.

V súčasnom období sa na opis obsahu súborov čoraz častejšie používa identifikátor media type (MIME type, content type), ktorého hodnota pozostáva minimálne z dvoch častí (napr. application/octet-stream). V obecnom prípade môže byť tých častí viac (napr. image/tiff/5.0). Hodnoty media type celosvetovo definuje Internet Assigned Numbers Authority (IANA) Autorita pre prideľovanie čísel na Internete (http://www.iana.org/).

K dnešnému dňu sú na najvyššej úrovni identifikácie media type registrované tieto mená:

application, audio, example, image, message, model,multipart, text, video.

V registroch IANA (http://www.iana.org/assignments/media-types/media-types.xhtml) boli ku dňu 27. 9. 2016 evidované tieto počty media type:

application (1190)
audio (144)
example (56)
image (56)
message (21)

model (22)
multipart (15)
text (72)
video (78)

Osobitný problém pri práci so súbormi predstavuje identifikácia formátu súboru. V bežnej praxi sa formáty súborov identifikujú:

podľa prípony súboru
podľa magických čísel
podľa explicitných metadát (Rozšírené atribúty súborov v niektorých operačných systémoch, MIME type, ...)

V náročnejších aplikáciách je takáto identifikácia formátu súboru nespoľahlivá. V takýchto prípadoch sa na identifikáciu formátu súboru používajú špeciálne SW produkty, ktoré dokážu spoľahlivo identifikovať obsah súboru. Napr. služba PRONOM poskytuje na identifikáciu obsahu súborov nástroj DROID (Digital Record Object Identification).

V prípade video formátov a audio formátov majú osobitný význam kodeky.

5. Formátová stratégia Centrálneho dátového archívu

Formáty súborov (digitálne formáty) sú jednou zo základných odborných tém v rámci dlhodobého uchovávania údajov. CDA ako LTP archív je s touto problematikou dennodenne konfrontovaný a venuje jej systematickú pozornosť.

Je veľmi ťažké, takmer nemožné, aby jednotlivci, resp. jednotlivé subjekty, ktoré prevádzkujú LTP archívy, sledovali, resp. evidovali problematiku vývoja jednotlivých formátov súborov v celosvetovom meradle. Vo svete existuje a existovalo viac projektov zameraných na tento účel. Napr.:

PRONOM (http://www.nationalarchives.gov.uk/PRONOM/Default.aspx)
Unified Digital Formats Registry (UDFR) (http://udfr.org/)
Global Digital Format Registry (GDFR) (http://library.harvard.edu/preservation/digital-preservation_gdfr.html)

CDA využíva vo svojej činnosti výsledky projektu PRONOM. Projekty UDFR a GDFR už nie sú aktívne.

V rámci LTP archívu je užitočné udržiavať samostatnú entitu, ktorá eviduje všetky formáty súborov, v ktorých sú uložené súbory v archíve. Inými slovami povedané, obsahuje evidenciu formátov súborov, ktoré sú pre daný archív povolené. V prípade CDA túto funkciu plní Formátová databáza CDA.

Vo Formátovej databáze CDA sú pre každý povolený formát súboru, okrem iného, uvedené tieto údaje:

Jednoznačný identifikátor formátu súboru – Unikátny identifikátor formátu registrovaný službou PRONOM (PUID) (napr. fmt/645) alebo jednoznačný identifikátor formátu pridelený CDA proprietárnemu formátu súboru, ktorý sa nenachádza v databáze PRONOM z rôznych dôvodov (napr. cda/102)
MIME Type – Opis obsahu súborov podľa platných štandardov MIME [8]
Názov formátu – Pomenovanie formátu podľa štandardu formátu alebo normy formátu, alebo podľa iného dokumentu popisujúceho formát (napr. Tagged Image File Format for Elektronic Photography – TIFF/EP)
Risks – Akékoľvek riziko spojené s daným formátom
Supported until – Dátum, dokedy bol alebo dokedy bude daný formát podporovaný
Dátum poslednej modifikácie formátu – Dátum, kedy bol formát naposledy modifikovaný

Zo strategického hľadiska LTP archívov je dôležité, aby počet formátov súborov v ktorých sú uložené súbory v archíve bol primeraný poslaniu archívu, pokiaľ sa dá minimálne možný, aby to boli formáty štandardizované, resp. formáty, kde je zaručená ich kontinuita a vývoj. Samozrejme, zabezpečiť takúto požiadavku nie je vôbec ľahké a v niektorých prípadoch ani možné. Každá zmena formátu súboru, resp. jeho zánik, predstavuje pre LTP archívy minimálne riziko včasného rozpoznania tejto skutočnosti a hľadania riešenia ako existujúce súbory v tomto formáte konvertovať do iných formátov súborov tak, aby boli pre používateľa stále dostupné.

Formáty, ktoré akceptuje CDA k 31. 5. 2016 sú uvedené na Obr. č. 1: Formáty, ktoré akceptuje CDA k 31. 5. 2016.

PUID	MIME TYPE	VALIDATOR	Poznámka	Typ validatora	Identifikator
cda/101	application/vnd.cda.container.x-dpx/undef	cda.plugins.format.octet-stream		JHove validator [1]	Proprietarny
cda/102	application/vnd.cda.container.pusr/undef	cda.plugins.format.octet-stream		JHove validator	Proprietarny
fmt/1	audio/x-wav/undef	cda.plugins.format.wav		JHove validator	DROID
fmt/2	audio/x-wav/undef	cda.plugins.format.wav		JHove validator	DROID
fmt/6	audio/x-wav/undef	cda.plugins.format.wav		JHove validator	DROID
fmt/19	application/pdf/1.5, application/pdf/undef	cda.plugins.format.pdf		PDFiTextValidationPlugin	DROID
fmt/41	image/jpeg/undef	cda.plugins.format.jpeg		JHove validator	DROID
fmt/43	image/jpeg/1.01	cda.plugins.format.jpeg		JHove validator	DROID
fmt/44	image/jpeg/1.02	cda.plugins.format.jpeg		JHove validator	DROID
fmt/94	model/vrml/undef	cda.plugins.format.vrml	Starý validátor	Chisel	DROID
fmt/101	text/xml/undef	cda.plugins.format.xml		XMLWellFormednessValidationPlugin	DROID
fmt/142	audio/x-wav/undef	cda.plugins.format.wav		JHove validator	DROID
fmt/156	image/tiff/5.0, image/tiff/6.0	cda.plugins.format.tiff		JHove validator	DROID
fmt/193	application/octet-stream/undef	cda.plugins.format.octet-stream		JHove validator	DROID
fmt/208	application/octet-stream/undef	cda.plugins.format.octet-stream		JHove validator	DROID
fmt/353	image/tiff/5.0, image/tiff/6.0	cda.plugins.format.tiff		JHove validator	DROID
fmt/355	application/rtf/undef	cda.plugins.format.octet-stream		JHove validator	DROID
fmt/436	image/tiff/6.0	cda.plugins.format.tiff		JHove validator	DROID
fmt/541	application/octet-stream/undef	cda.plugins.format.octet-stream		JHove validator	DROID
fmt/645	image/jpeg/undef	cda.plugins.format.jpeg		JHove validator	DROID
fmt/703	audio/x-wav/undef	cda.plugins.format.wav		JHove validator	DROID
fmt/704	audio/x-wav/undef	cda.plugins.format.wav		JHove validator	DROID
x-fmt/111	text/plain/undef (UTF-8 bez BOM)	cda.plugins.format.text		JHove validator (UTF-8)	Enca
x-fmt/387	image/tiff/6.0	cda.plugins.format.tiff		JHove validator	DROID
x-fmt/391	image/jpeg/undef	cda.plugins.format.jpeg		JHove validator	DROID
x-fmt/392	image/jp2/undef	cda.plugins.format.jp2		JHove validator	DROID
fmt/569	video/x-matroska		Budúcnosť		DROID??
fmt/569	audio/x-matroska		Budúcnosť		DROID??

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

Obr. č. 1: Formáty, ktoré akceptuje CDA k 31. 5. 2016

Akceptované formáty sú výsledkom riešenia digitalizačných, infraštruktúrnych a dopytových projektov OPIS PO2 [2]. Vzhľadom na to, že niektoré digitalizačné projekty pokračujú aj v období udržateľnosti projektu a s prihliadnutím na to, že nie všetky dáta, ktoré vznikli počas riešenia digitalizačných a dopytových projektov sú uložené v CDA, nemožno považovať množinu akceptovaných formátov k 31. 5. 2016 za definitívnu.

Z pohľadu CDA treba doriešiť minimálne dva problémy. Aké formáty a kodeky povoliť pre uchovávanie audiovizuálnych dokumentov a 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.

V prípade audiovizuálnych dokumentov treba navrhnúť jeden alebo viac kontajnerových formátov, video formáty, audio formáty, video kodeky, audio kodeky, prípadne textové kodeky. Ako perspektívny kontajnerový formát pre audiovizuálne dokumenty sa javí formát Matroska (https://www.matroska.org/).

V prípade proprietárnych formátov, ktoré CDA nepovažuje za perspektívne a pri vklade by nemali byť súbory v tomto formáte komplexne testované a potom dlhodobo udržiavané, sú dve možnosti riešenia. Povoliť takéto formáty alebo vytvoriť nový CDA proprietárny kontajnerový formát, v ktorom sa bude vytvárať obálka pre súbory v neperspektívnych proprietárnych formátoch. V oboch prípadoch sa dostávame do rozporu s normou STN ISO 14721:2014 (OAIS) [3] (Napr.: Z časti 2.2.1 „Aby mohol byť tento Informačný objekt úspešne uchovávaný, je kritické pre OAIS, aby bol schopný jasne identifikovať a chápať Dátový objekt ako aj jeho Prezentačnú informáciu. V prípade digitálnej informácie to znamená, že OAIS musí jasne identifikovať bity ako aj Prezentačnú informáciu k nim sa vzťahujúcu. Táto požadovaná transparentnosť až po úroveň bitov je špecialitou uchovávania digitálnych informácií, a je v rozpore s objektovo orientovanými konceptmi, ktoré sa práve tieto implementačné detaily pokúšajú schovať. Toto predstavuje zásadnú výzvu pri uchovávaní digitálnych informácií“. *Prezentačná informácia a identifikácia „až po úroveň bitov“ znamená v prostredí CDA formátovú indentifikáciu a validáciu*). Ideálne by bolo, takéto formáty v CDA nepovoliť.

Formátová databáza CDA sa prioritne využíva v procese vkladu (ingescie) SIP balíkov do archívu, presnejšie v procese tvorby AIP balíkov z dodaných SIP balíkov. Pre každý relevantný súbor z vkladaného SIP balíka sa vykoná identifikácia formátu (DROID, Enca, proprietárna identifikácia) a následne jeho validácia príslušným validátorom.

Formátová databáza CDA je základným nástrojom na identifikáciu a riešenie formátových rizík. V pravidelných (mesačných) intervaloch sa synchronizuje s databázou PRONOM pričom sa aktualizujú hodnoty premenných *Risks, Supported until*. Hodnota premenných *Risks, Supported until* sa môže v odôvodnených prípadoch meniť aj prostredníctvom užívateľského rozhrania. Analýza hodnôt premenných *Risks, Supported until* slúži ako podklad pre rozhodovanie o prípadnej formátovej konverzii alebo o iných opatreniach.

Formátovú konverziu môže vykonávať:

CDA po dohode s vkladateľom (člen určeného spoločenstva)vkladateľ pomocou vlastných technických a personálnych prostriedkovtretia strana, na ktorú bude táto úloha delegovaná

Konverzia formátu neznamená nahradenie starého formátu novým, ale doplnenie nového formátu do Formátovej databázy CDA.

Formátová konverzia prináša so sebou aj identifikáciu rizikových AIP balíkov.

V prípade LTP archívov si treba uvedomiť, že v dlhodobom úložisku nie je prípustné uložené objekty ani modifikovať ani vymazávať. Súbor uložený v AIP v starom formáte sa konvertuje do súboru v novom formáte a potom sa pridá do AIP. V LTP archívoch slúži na takúto operáciu, v súlade s normou STN ISO 14721:2014 [3], proces vytvárania Verzií AIP balíkov.

6. Záver

CDA je koncipovaný a funguje ako dlhodobé dôveryhodné úložisko digitálneho obsahu.

CDA bude aj naďalej v maximálnej možnej miere uplatňovať otvorené formáty súborov, vrátane kontajnerových, a otvorené kodeky (audio, video, text).

CDA sa bude aj naďalej orientovať na také formáty súborov, ktoré budú mať k dispozícii kvalitné validátory. Pre CDA sú zaujímavé napr. očakávané výsledky projektu PREFORMA (http://www.preforma-project.eu/index.html).

CDA bude v krátkej dobe akceptovať jeden kontajnerový formát pre audiovizuálne dokumenty, vrátane video formátov, audio formátov, video kodekov, audio kodekov, prípadne textových kodekov.

CDA sa pokúsi 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. Odporučí určenému spoločenstvu vhodnejšie akceptované alebo akceptovateľné formáty.

CDA sa bude aj naďalej v pravidelných intervaloch zaoberať identifikáciou a riešením formátových rizík.

CDA sa bude aj naďalej usilovať aby počet formátov súborov, v ktorých sú uložené súbory v archíve, bol primeraný poslaniu archívu a pokiaľ sa dá, aby bol minimálne možný.

CDA bude aj naďalej sledovať vývoj v oblasti formátov, kodekov 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.

Použité skratky

AIP – Archival Information Package (Archívny informačný balík)
CDA – Centrálny dátový archív
DIP – Dissemination Information Package (Výberový informačný balík)
HW – Hadvér
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)
SW – Softvér
UKB – Univerzitná knižnica v Bratislave

Použitá literatúra

[1] CIGLAN, Ivan. Národný projekt Centrálny dátový archív. In: *ITlib.* 2012, č. 2, s 35-36. ISSN 1335-793X.

[2] *Operačný program Informatizácia spoločnosti – Prioritná os 2* [online]. [cit. 2016-10-10]. Dostupné z: http://www.opis.culture.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*. 2016, č. 2, s. 37-52. ISSN 1335-793X.

[5] SCHAEFER, Sibyl. Trustworthy Digital Preservation Repositories: an Introduction. In: *ITlib*. 2016, č. 2, s. 53-55. ISSN 1335-793X.

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

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

[8] *MIME. Multipurpose Internet Mail Extension* [Viacúčelové rozšírenie internetovej pošty] [online]. [cit. 2016-10-10]. Dostupné z: https://en.wikipedia.org/wiki/MIME

Formáty a LTP v Európe

Bibiána Žigová, Univerzitná knižnica v Bratislave

Abstrakt:
Článok analyzuje stratégiu uchovávania digitálneho obsahu v rôznych formátoch z hľadiska LTP (Long Term Preservation). Obsahuje informácie a porovnanie používaných formátov v krajinách EU. Informuje o stave LTP v Európe.

Kľúčové slová:
digitálne objekty, digitálny obsah, dátové úložisko, dlhodobé uchovávanie digitalizovaného kultúrneho dedičstva, LTP, ochrana digitálneho obsahu, formáty.

Úvod

Stratégia uchovávania digitálneho obsahu je kľúčová a túto otázku riešia takmer na celom svete. Uchovávanie zahŕňa širokú škálu úkonov určených na zabezpečenie využiteľnosti obsahu a jeho ochranu pred fyzickým zničením, stratou, nefunkčnosťou média, na ktorom je uložený.

Digitalizácia je finančne a časovo náročný proces a je preto nevyhnutné pracovať efektívne a zabezpečiť trvalé uchovanie digitalizovaného obsahu pre budúce generácie. Nástrojom pre tento proces je štandardizácia jednotlivých krokov digitalizácie od skenovania až po vklad do elektronického archívu.

V súčasnosti je digitálny obsah uchovávaný na magnetických médiách (LTO,JAG), optických médiách (CD,DVD), na diskových poliach serverov, či osobných pracovných staníc a na prenosných externých diskoch. Stratégia uchovávania sa v zásade týka oblasti médií, na ktorých sa obsah uchováva a oblasti formátov, v ktorých sa dáta uchovávajú, prípadne sprístupňujú. V tomto príspevku sa budem venovať primárne formátom pre dlhodobé uchovávanie dát.

Riešenie oblasti formátu dát je postavené na dvoch základných pilieroch: formátová ochrana a bitová ochrana. Zmena štandardov a vývoj technológií podmieňuje konver-

ziu súborov do formátov, ktoré sú v danej dobe známe, čitateľné a tak je obsah dostupný pre používateľov.

Čo je to formát

Formát je sada sémantických a syntaktických pravidiel pre mapovanie medzi abstraktným informáciami a ich digitálnou podobou. Formát, často nazývaný aj súborový formát, typ súboru, je konkrétny spôsob kódovania informácií s cieľom ich uchovania v počítačovom súbore. (https://sk.wikipedia.org/wiki).

Aké formáty poznáme

Formáty súborov sú navrhnuté väčšinou tak, že uchovávajú konkrétne typy dát. Napríklad JPEG na obrázky, GIF na obrázky a jednoduché animácie, HTML pre zdrojový kód. Informačno-komunikačná technika sa pozerá na súbory ako na tok dát, na nuly a jednotky, je nevyhnutná metóda na určenie formátu konkrétneho súboru. Určenie sa môže diať podľa prípony súboru, podľa špecifickej sady bajtových identifikátorov vnútri súboru, podľa metadát, podľa mime type.

Registre formátov

Pre dlhodobé uchovávanie je dôležité identifikovať formát, v ktorom je zdigitalizovaný obsah uložený, jeho špecifiká a medzinárodne dohodnuté označenie. Uvádzam príklad využívaných registrov.

- **IANA**

  Organizácia IANA spravuje zoznam typov internetových médií (Internet Media Type) „Mime Type“. (www.iana.org). Zoznam spravuje aj UDFR a PRONOM.

- **UDFR**

  Unified Digital Format Registry (UDFR) sémantický register formátov pre digitálne uchovávanie. UDFR bol spoľahlivý, verejne prístupný register znalostnej bázy formátov súborov informácií, vhodných pre uchovávanie digitálneho obsahu. UDFR sa snažil „zjednotiť“ funkciu a držanie dvoch existujúcich registrov, PRONOM a GDFR (Global Digital Format Register) vo forme open source platformy. UDFR bol vyvinutý Univerzitou v Kalifornii (University of California) – California Curation Center (UC3) v rámci Kalifornskej digitálnej knižnice (California Digital Library), pod patronátom Kongresovej knižnice (Library of Congress). Tvoril súčasť programu pre digitálne uchovávanie informácií (National Digital Information Infrastructure Preservation Program). Činnosť UDFR bola ukončená 15. apríla 2016 a odporúča sa používať register PRONOM.
  (http://www.udfr.org/)

- **PRONOM**

  Iniciatívou National Archives UK je register PRONOM (http://www.nationalarchives.gov.uk/pronom/). Databáza obsahuje súbor formátov, softvérov, dodávateľov, jednotlivé typy, skratky, verzie formátov (Mime type, ID, PUID). Uvádza aj mieru risku pri zastaraných formátoch. Súčasťou služieb Národného archívu UK je DROID (Digital Record Object Identification) – systém automaticky identifikujúci formát. Tento systém je k dispozícii zdarma na stránkach NA UK (http://www.nationalarchives.gov.uk/pronom/).

Formáty vhodné pre LTP

Najkomplexnejšie procesy dlhodobého úložiska súvisia s potrebou zabezpečenia dlhodobého uchovania obsahu na úrovni formátov. Ide o zabezpečenie zrozumiteľnosti obsahu v budúcnosti. Realizuje sa niekoľkými spôsobmi:

- Migrácia – dokument je v prípade zastaraného formátu skonvertovaný do nástupníckeho formátu. Pôvodný formát je vymazaný a ďalej sa neuchováva.
- Normalizácia – dokument je skonvertovaný do alternatívneho formátu, ktorý je považovaný za stabilný. Pôvodný formát je uchovaný tiež.
- Emulácia – v prípade zastarania formátu vznikne výpočtové prostredie, v ktorom je možné formát kedykoľvek interpretovať. Prostredie musí byť prispôsobené IKT technológiám v danom čase.

Dôvodom pre migráciu dát z pohľadu OAIS je:

- Starnutie médií – časom klesá spoľahlivosť a životnosť médií.
- Väčšia efektívnosť nových typov médií – väčšia kapacita médií za nižšiu obstarávaciu cenu.
- Nové požiadavky používateľov úložiska – nové technológie.
- Evolúcia softvéru – rizikové formáty a vývoj nových formátov.

Cieľom migrácie dát podľa OAIS je dlhodobé uchovanie dát a ich ochrana pred neželanými vplyvmi nasledujúcimi procesmi:

- Obnovenie – obnovuje sa na novšie alebo iné médium, dáta zostávajú nezmenené.
- Replikácia – obnovuje sa na médium rovnakého typu, dáta zostávajú nezmenené.
- Prebalenie – dáta sa prebalia a zmení sa „packaging information“ v AIP balíku.
- Transformácia – mení sa formát dát, vytvorí sa nová verzia AIP balíka, vzniká nové AIP ID.
- Nová edícia – vylepšuje sa AIP balík, napr. doplnením dát do AIP balíka, vzniká nové AIP ID.

Veľké archívy plné papierových dokumentov sa veľmi ťažko organizujú, nevyhnutný je veľký priestor s vyhovujúcou statikou, špeciálne podmienky uskladnenia – vlhkosť, teplota, svetlosť priestoru, navyše pri vzácnych dokumentoch aj fyzická ochrana. Digitálny archív má nepomerne viac výhod ako klasický. Dokumenty sú uchované bez ohľadu na stav papiera, sú prístupné väčšiemu auditóriu, dostupné prakticky online z domu, ľahšie sa v nich vyhľadáva a archív zaberá menej miesta.

Výsledkom digitalizácie sú dáta, ktoré sú síce v štandardných formátoch súborov, ktorých producentmi sú softvérové spoločnosti. Avšak ani to nezaručí, že formáty sú vhodné pre dlhodobé uchovanie. Digitálne dáta musia totiž pri vklade do dlhodobého archívu prejsť množstvom validačných krokov, ktoré zabezpečia, že daný súbor bude čitateľný, použiteľný v dlhodobom horizonte.

Vo svete sa najčastejšie stretávame s uchovávaním digitálnych dokumentov vo formáte PDF/A pre texty. Je to verzia formátu PDF, ktorá vypúšťa niektoré funkcie, napr. šifrovanie, zvukový a video obsah a prepojenie na externé zdroje. Pre obrázky je to TIFF a JPEG2000. Pre videá DPX. Veľké národné archívy si kladú ako jednu z podmienok vklad dokumentov v niekoľkých formátoch tak, aby vedeli zabezpečiť dlhodobé uchovanie a prípadnú formátovú konverziu.

Situácia v Európe

- **Poľsko**
  V roku 2007 začala v Poľsku pôsobiť Národná digitálna knižnica „POLONA“ (https://polona.pl). Bola vytvorená pre všetky knižnice a pre všetkých internetových používateľov s poslaním rozšíriť široký a jednoduchý prístup k digitálnym zbierkam Poľskej národnej knižnice, vrátane literatúry a vedeckých materiálov, historických dokumentov, časopisov, grafiky, fotografie, notových zápisov a máp. V roku 2014 bola pre verejnosť otvorená Digitálna požičovňa vedeckých publikácií „Academica“ (https://academica.edu.pl), ktorá plní klasickú funkciu medziknižničnej výpožičnej služby, avšak len digitálnych diel.

- **Česká Republika**
  Národní knihovna ČR v Prahe sprístupnila portál Národní digitální knihovna ČR, ktorý využíva systém „Kramerius“ (http://www.ndk.cz) digitalizuje, sprístupňuje a dlhodobo uchováva zdigitalizované kultúrne dedičstvo ČR. Systém Kramerius je postavený na open source riešení Fedora. Pre všetky tieto činnosti je nevyhnutné stanoviť a dodržiavať štandardy vytvárania dát a metadát.

Národní digitální knihovna využíva systém persistentných identifikátorov, založených na štandarde URN:NBN (https://resolver.nkp.cz). „ČIDLO“ (CZIDLO – CZech IDentification and LOcalization tool) je služba pre knižnice a pamäťové a fondové inštitúcie z oblasti českého kultúrneho dedičstva pre potreby trvalej identifikácie digitálnych dokumentov alebo ako prostriedok pre zaistenie dôveryhodnosti citačnej praxe (overovanie autenticity citovaných dokumentov).

Kolegovia v NK ČR v Prahe prieskumom v oblasti formátov súborov elektronických publikácií zistili, že preferované formáty pre dlhodobú ochranu sú EPUB 2.0.1 a PDF/A-1b. Zároveň testovali aj validátori a ako najvhodnejšie sa javia pri EPUB softvérový modul Epubcheck, pre PDF/A-1b bol potvrdený Apache PDFBox. LTP úložisko NK ČR v rámci kompletného workflow pre spracovanie a uchovanie elektronických publikácií ukladá elektronické publikácie vo vopred určených formátoch s validnými metadátami podľa vytvoreného štandardu.

Moravská zemská knihovna v Brne úzko spolupracuje s Národní knihovnou ČR v Prahe a podobne sprístupňuje Digitální knihovnu v systéme Kramerius. Digitalizuje, sprístupňuje a dlhodobo uchováva zdigitalizované regionálne diela.

- **Platforma V4**

Medzi knižnicami V4 prebieha už tradične spolupráca na veľmi dobrej úrovni. Prebieha mnoho aktivít a vynakladá sa značné úsilie v oblasti získavania, ochrany digitálnych zdrojov. Tieto metodické a praktické znalosti majú veľkú hodnotu. Spolupráca v rámci platformy V4 sa týka najmä nasledujúcich oblastí:

- certifikácia dlhodobých úložísk podľa medzinárodných noriem ISO 16363, ISO 14721, ISO 27001,
- harmonizácia legislatívy v oblasti e-born dokumentov, autorského práva online dokumetov,
- metodika uchovávania zdigitalizovaného kultúrneho dedičstva,
- webharvesting.

- **PREFORMA**

Iniciatíva PREFORMA (http://www.preforma-project.eu) je projekt „PREservation FORMAts“ uchovávania formátov pre elektronické archívy. Projekt začal 1.1.2014 a spolutvorcom je Európska komisia a jej program „FP7-ICT Programme“ (http://cordis.europa.eu/fp7/ict/home_en.html). Cieľom projektu je vyriešenie problému zavádzania kvalitných štandardizovaných formátov súborov pre uchovanie obsahu dát v dlhodobom horizonte. Hlavným cieľom je poskytnúť pa-

mäťovým inštitúciám plnú kontrolu nad priebehom validácie súborov pri vklade do archívu.

Iniciatíva PREFORMA poskytuje priestor expertom, inštitúciám a iným projektom diskutovať, nájsť riešenia a poskytnúť ich verejnosti. Cieľom iniciatívy je spojiť pamäťové a fondové inštitúcie s vývojármi, softvérovými spoločnosťami a normalizačnými agentúrami, poskytnúť všetkým zainteresovaným stranám spätnú väzbu. Výsledkom by mal byť súbor formátov s adekvátnymi validátormi, vhodnými pre dlhodobú archiváciu. Členmi iniciatívy sú pamäťové a fondové inštitúcie, archívy, knižnice, softvérové firmy z Belgicka, Švédska, Holandska, Nemecka, Rakúska, Grécka, Talianska, Španielska, Estónska, Veľkej Británie, Slovenska. CDA je členom iniciatívy PREFORMA od júla 2016.

- **UNESCO**

Charta o zachovaní digitálneho dedičstva, prijatá na 32. Zasadnutí Generálnej konferencie UNESCO v roku 2003 v časti Digitálne dedičstvo ako spoločné dedičstvo položila základy digitalizácie hmotného a nehmotného kultúrneho dedičstva. Následne Vancouverská deklarácia UNESCO z roku 2012 „Pamäť sveta v digitálnom veku: digitalizácia a uchovávanie“ stanovila rámec medzinárodnej spolupráce pri uchovávaní zdigitalizovaného kultúrneho dedičstva sveta. Konzorcium pamäťových a fondových inštitúcií stanovené deklaráciou na prácu v rokoch 2013 – 2015 usporiadalo workshopy, konferencie a panelové diskusie, kde odborníci z oblasti knihovníctva, archívnictva, múzeí, štátnej správy a komerčných firiem diskutovali o urgentnej potrebe uchovávania digitálnych informácií pre budúcnosť. Jednou z tém diskusií bola problematika otvorených formátov, štandardizácia formátov, binárna ochrana dát a ochrana konzistentnosti dát. Diskusie mali za cieľ nájsť riešenia a ponúknuť ich odbornej verejnosti. Na základe posledných zistení sú softvérové riešenia dlhodobej ochrany digitálnych dát veľmi drahé a preto takmer nedostupné pre mnohé inštitúcie. Zároveň open source riešenia nie sú dostatočné, potreby kurátorov a používateľov sú nad možnosti týchto riešení.

Diskusie konzorcia sa týkali aj skúmania dopytu po digitalizovaných dátach. Skúmalo sa, kto aké dáta využíva, v akej forme ich potrebuje. Na základe posledných zistení je využívanie väčšiny digitálnych dát spoplatnené, čo obmedzuje prístup k nim.

Ďalším problémom, ktorý sa riešil, je kvalita digitalizovaných dát a ich dôveryhodnosť. Na základe zistení mnoho digitálnych informácií nemá základnú štruktúru,

nemá, alebo má len minimum metadát, čo znemožňuje vyhľadávanie, prácu s nimi. Sú de facto nedostupné.

Odborníci z oblasti digitalizácie upozornili na fakt, že sa stále mylne interpretuje digitalizácia a uchovávanie digitálneho obsahu. Nevyhnutná je osveta aj v tomto smere. Digitalizácia zahŕňa nielen transformáciu obsahu do digitálnej formy. Zahŕňa mnoho aktivít, od akvizície, identifikácie, čistenia, reštaurovania, katalogizácie, vytvorenia metadát až po uloženie digitalizovaného obsahu na špeciálnom médiu (magnetickom, optickom) do trezora či archívu.

Uchovávanie to nie je len fyzické zachovanie kópie, kópií v trezoroch. Ide hlavne o možnosť využiť digitalizované dáta v budúcnosti, za podmienky zachovania úplnej integrity dát. (http://www.unesco.sk/O-UNESCO)

Pre pamäťové a fondové inštitúcie je kľúčové stanovenie stratégie digitalizácie a uchovávania digitalizovaných dát. Stratégia by mala obsahovať nielen obsah dát, ale hlavne analýzu efektivity nákladov, prípadné výnosy, mala by identifikovať potenciálnych donorov, možnosti finančnej podpory vlády daného štátu.

- **Dánsko**

Dánsky národný archív (https://www.sa.dk/en/about-us/danish-national-archives) uchováva a sprístupňuje zo zákona v papierovej aj digitálnej forme historické dokumenty Dánska, dáta štátnej správy a justície. Zákon jasne stanovuje formu, formáty, metodiku tvorby SIP balíka, dlhodobé uchovanie až po sprístupňovanie. Zákon stanovuje nasledujúce možné formáty v SIP balíku: TIFF, MP3, MPEG-2, MPEG-4, JPEG2000, GML, WAVE. Pre zaujímavosť dodávam, že v SIP balíky do Dánskeho národného archívu môžu byť dodané na CD-R, DVD-R prípadne na USB médiu. V jednom SIP balíku môžu byť zastúpené všetky spomenuté prípustné formáty.

- **Švédsko**

Švédsky národný archív (https://riksarkivet.se) uchováva a sprístupňuje zo zákona všetky dokumenty štátnych inštitúcií a historické dokumenty kultúrneho dedičstva vo Švédsku. Metodicky usmerňuje, koordinuje vedu a výskum v oblasti formátov pre dlhodobé uchovávanie v rámci iniciatívy PREFORMA.

- **Belgicko**

V Belgicku vznikla pozoruhodná organizácia „PACKED“ – centrum analýz a expertíz v oblasti uchovania digitálneho dedičstva (http://www.packed.be/en/). Je to

platforma pre spoluprácu organizácií archivujúcich a dlhodobo uchovávajúcich audiovizuálne diela. Cieľom práce organizácie je zlepšovať kvalitu a efektivitu pri digitalizácii a archivovaní všetkých typov dokumentov, najviac však audiovizuálnych, mediálnych diel. PACKED úzko spolupracuje s inštitúciami spojenými do projektu „EUROPEANA“ (www.europeana.eu).

- **EUROPEANA**

Europeana je digitálna knižnica, ktorá umožňuje širokej verejnosti prezerať si digitálne zdroje z európskych múzeí, knižníc, archívov a audiovizuálnych zbierok. Portál Europeana (www.europeana.eu) bol spustený 20. novembra 2008 Viviane Redingovou, eurokomisárkou pre informačnú spoločnosť a médiá. Projekt je realizovaný centrálnym tímom, sídliacim v národnej knižnici Holandska – Koninklijke Bibliotheek. Europeana je internetový portál, ktorý funguje ako rozhranie k miliónom kníh, obrazov, filmov, múzejných predmetov a archívnych záznamov v celej Európe. Do Europeanei prispieva viac ako 2000 inštitúcií. (Louvre či British Library až po regionálne archívy a miestne múzeá Európskej únie). Umožňuje prístup k rôznym typom obsahu (digitálne objekty nie sú uložené na centrálnom počítači, ale zostávajú v jednotlivých inštitúciách). Europeana zhromažďuje kontextové informácie (metadáta) o položkách, vrátane malého obrázku. Pre prístup k celému obsahu, musíte prejsť na pôvodný web, ktorý obsah drží.

- **Rakúsko**

Dlhodobej ochrane digitalizovaných dokumentov sa venuje v Rakúsku viac inštitúcií. Z môjho pohľadu je zaujímavá práca digitálnej mediatéky „Die Österreichische Mediathek“ (http://www.mediathek.at/). Nielen preto, že digitalizujú a uchovávajú rakúske audiovizuálne dokumenty, ale preto, že ponúkajú tieto služby aj širokej verejnosti. Od roku 2000 majú k dispozícii digitalizačný systém (DVA Profession), ktorý vyvinuli na mieru v spolupráci so softvérovými spoločnosťami. Systém ich digitalizácie pozostáva z katalogizácie, kde sa produkujú popisné metadáta a katalogizačný záznam s jedinečnou signatúrou do databázy DABIS. Následne prichádza fyzické, mechanické čistenie pôvodného média, reštaurátorské práce, napríklad zlepenie roztrhnutej pásky audio kazety. Proces pokračuje produkciou technických metadát o type média, type pásky, platne, rýchlosti prehrávania, o výrobcovi, prehrávači. Samotná digitalizácia je automatický workflow pozostávajúci z validácie formátu, kontroly celistvosti, kontroly checksúm MD5, výroby náhľadu. Digitalizácia sa končí uložením do dlhodobého úložiska typu RAID v dvoch kópiách. Prvá kópia je archívny master, druhá je náhľadová, určená pre prezeranie a bádanie.

Pre audio využíva rakúska mediatéka:

- Master – Broadcast Wave v kvalite 96 kHz/24 Bit
- Náhľad – MP3 v bitrate 128 kbit/s

Pre video využíva rakúska mediatéka:

- Master – video codec FFV1, audio codec PCM bez komprimácie, container AVI
- Náhľad – MPEG-2

- **Nemecko**

Priekopníkom v Nemecku je nemecká národná knižnica „Die Deutsche Nationalbibliothek“ (http://www.dnb.de/DE/Home/home_node.html). Už od roku 2004 zastrešuje niekoľko projektov s témou digitalizácie a LTP. Najznámejší je projekt „KOPAL“, ktorého výsledkom bol formát pre elektronické zdroje „LMER“, softvér pre ingest a disemináciu dát „koLibRi“, či projekt „AREDO“ pre LTP. Významným projektom je „Nestor“- centrálny systém pre metodiku, certifikácie, hodnotenie dlhodobých dôveryhodných úložísk.

Archívny systém nemeckej národnej knižnice bol certifikovaný ako dôveryhodné úložisko pečaťou kvality „Data Seal of Approval“. Vo svojich politikách uvádza ingest niekoľkých fixných formátov: PDF, TIFF, JPEG, PS, EPUB.

- **Veľká Británia**

Národný archív Veľkej Británie (http://www.nationalarchives.gov.uk) plní funkcie štátneho archívu. Metodicky zastrešujú digitálne kurátorstvo. Uchováva len stabilné formáty, vhodné pre dlhodobé uchovanie. „The National Archives“ preferujú ingest týchto formátov: TIFF, JPEG, JPEG2000, PDF. Akceptujú aj ingest iných formátov, ale neuchovávajú proprietárne formáty, otvorené štandardy pre dokumenty, dáta v ASCII, Unicode.

http://www.nationalarchives.gov.uk/about/commercial-opportunities/digitisation-services/our-digitisation-services/

- **Slovensko**

Na Slovensku prebiehala a stále prebieha digitalizácia kultúrneho dedičstva v jednotlivých PFI samostatne v období rokov 2011 – 2015. Cieľom bolo vybudovať Slovenskú digitálnu knižnicu, sieť špecializovaných digitalizačných pracovísk PFI, centrálny archív pre dlhodobé uchovávanie, webharvesting, webarchiving, centrálny portál pre sprístupňovanie zdigitalizovaného obsahu.

Samostatnou kapitolou boli legislatívne opatrenia v oblasti autorskoprávnej ochrany diel. Zohľadnené boli medzinárodné normy a odporúčania OAIS, LTP, METS, XFDU, DCMES, PREMIS, ako aj skúsenosti a know-how z existujúcich systémov pre dlhodobé uchovávanie digitálneho obsahu.

Sprostredkovateľský orgán rezortu kultúry vydal pre projekty pamäťových a fondových inštitúcií v rámci OPIS PO2 (Operačný program informatizácia spoločnosti, Prioritná os 2 – Rozvoj pamäťových a fondových inštitúcií a obnova ich národnej infraštruktúry. Operačný program je spolufinancovaný zo zdrojov Európskeho fondu regionálneho rozvoja. http://www.opis.gov.sk) niekoľko metodických manuálov. Jedným z nich je metodický manuál č.17 pre zabezpečenie dlhodobej archivácie konvertovaných objektov. Manuál vytvoril rámec pre pamäťové a fondové inštitúcie (PFI) v oblasti digitalizácie kultúrneho dedičstva Slovenska. Stanovil východiská a princípy pre metadáta, formáty, registre, číselníky, SIP/AIP/DIP balíky. Referenčným zdrojom sú normy ISO 16363, ISO 14721, Centrálny dátový archív Univerzitnej knižnice prevzal a pripravil slovenské verzie STN. Referenčným zdrojom pre formáty je register PRONOM.

- **SLOVAKIANA**

Slovensko a jeho kultúrne dedičstvo reprezentuje portál „SLOVAKIANA“ (https://www.slovakiana.sk/). Bol spustený v novembri 2015, ktorý pre odbornú aj laickú verejnosť sprístupňuje výsledky digitalizácie slovenského kultúrneho dedičstva. Portál je súčasťou siete európskych kultúrnych portálov. Je výsledkom projektu CAIR „Centrálna aplikačná infraštruktúra a registratúra“ Národného osvetového centra NOC
(www.nocka.sk) v rámci OPIS PO2.

- **Centrálny dátový archív (CDA)**

Cieľom projektu OPIS PO2 – CDA bolo vybudovať komplexný integrovaný systém dlhodobého uchovávania a ochrany digitálneho obsahu, jeho získavania, spracovania, ochrany a využitia. V rámci projektu sa vybudoval centrálny dátový archív, ktorý v zmysle predpísaných štandardov zabezpečuje uloženie kópií digitálnych objektov v najmenej dvoch geograficky oddelených lokalitách, vzdialených minimálne 50 km. Lokality sú vybavené príslušným hardvérovým i softvérovým vybavením a prostredníctvom dátových telekomunikačných sietí sú vzájomne prepojené. Súčasťou dátového archívu je aj pasívny sklad pamäťových médií (terciárna lokalita). Projekt sa ukončil v roku 2014 a v súčasnosti je CDA v prevádzke druhý rok. CDA je vybudovaný na platforme OAIS v súlade s medzinárodným normami ISO 16363, ISO 14721. Je držiteľom certifikátu Systému manažérstva

informačnej bezpečnosti podľa normy ISO/IEC 27001. K septembru 2016 má uložených 1,5 milióna SIP balíkov v celkovej veľkosti 3,5 PetaByte. (http://cda.kultury.sk).

Čo s formátmi v Centrálnom dátovom archíve

V rámci OPIS PO2 je celkovo 9 národných projektov pamäťových fondových inštitúcií z rezortu kultúry a 7 dopytových projektov. To naznačuje aj veľkú rôznorodosť dát, veľa formátov.

Obr. 1 Mrak reálnych formátov uložených v CDA – čím väčšie logo, tým vyšší počet SIP balíkov s daným formátom

CDA je budovaný ako dlhodobý dôveryhodný archív (a len ako archív), ktorý poskytuje komplexné služby vyhradené pre určené spoločenstvo. Archív samotný nesprístupňuje a „nevlastní" uložený obsah. Služby sa poskytujú na základe písomnej *Dohody o zverení obsahu na dlhodobú archiváciu* , ktorej základnou súčasťou je tzv. volumetria dohodnutých objemov a frekvencií vkladu a výberu dát a tzv. profil vkladateľa, ktorý obsahuje zoznam digitálnych formátov. Vkladateľ sa zaväzuje, že bude vkladať v rámci vytvoreného profilu digitálne objekty v niektorom z uvedených formátov. V tabuľke sa uvádzajú všetky a podľa možností len tie typy formátov, ktoré sa vyskytnú vo vkladných balíkoch a spracujú pod daným profilom. Vkladateľ berie na vedomie, že ak bude doručený súbor s iným formátom, vklad bude zamietnutý.

CDA PROFIL (UKB_01)			
PUID	**MIMEType**	**Názov formátu**	**Poznámka**
fmt/43	image/jpeg	JPEG File Interchange Format v1.01	.jpg
fmt/96	text/html	Hypertext Markup Language	.html, .htm
fmt/101	application/xml text/xml	Extensible Markup Language v1.0	.xml
fmt/353	image/tiff	Tagged Image File Format	.tif
x-fmt/111	text/plain	Plain Text UTF 8	.txt
x-fmt/392	image/jp2	JP2 (JPEG 2000 part 1)	.jp2

Pozn.: údaje v tabuľke sú prevzaté z registra PRONOM
Príklad: Zoznam formátov z profilu vkladateľa

Dohoda neobmedzuje formátovú pestrosť vkladov, pri bilaterálnom rokovaní sa však uprednostňujú formáty podporované v registri PRONOM, ale prípustné sú po dohode aj niektoré proprietárne formáty, ktoré sa podľa potreby zabalia (TAR, ZIP) alebo ukladajú ako bitové objekty.

Formáty, ktoré sa nenachádzajú v PRONOM databáze, a ak inštitúcia trvá na uložení, vtedy sa balia do „zip kontajnera“, resp. do špeciálneho kontajnera vytvoreného priamo pre inštitúciu – napr. PUR kontajner pre Pamiatkový úrad SR. Pre takéto sa musí vytvárať špeciálny validátor pre špeciálne potreby konkrétnej PFI.

K takýmto situáciám sa prikláňame ale až vo veľmi výnimočných situáciách, pretože archív je LTP a pri proprietárnych „exotických“ formátoch nevieme zaručiť formátovú konverziu a ochranu.

Transfer digitálnych objektov medzi CDA a určeným spoločenstvom sa realizuje výhradne formou kompletných „zbalených“ informačných balíkov (SIP, DIP) v odporúčanom formáte GNU TAR, prípadne ZIP s predpísanou vnútornou štruktúrou. Použitie hierarchického formátu BagIt sa v praxi z dôvodu intenzívnych prenosov na páskach neosvedčilo.

Ako sa ukázalo, odlaďovanie vkladu je iteračný proces, vyžadujúci najmä trpezlivosť a flexibilitu na oboch stranách. Vkladateľ má na začiatku pridelený testovací profil, cez ktorý sa opakovane odlaďujú jednotlivé balíky najskôr na priechodnosť cez kontrolné procesy. Po odladení sa nasadí tzv. ostrý profil, ktorý prepustí vkladaný balík s danou identifikáciou len raz. Metodika vkladania dát do Centrálneho dátového archívu je ošetrená zmluvnými vzťahmi s PFI, kde sú deklarované formálne a obsahové charakteristiky dát, ktoré sú predmetom vkladu. Vstup dát do CDA je tak možný vý-

hradne na základe uvedených formalizovaných štruktúr. Ich cieľom je komplexne metodicky zabezpečiť procesy vkladu resp. výberu dát z CDA ako aj procesy súvisiace s logistikou a kontrolou kvality.

Typ dát	Akceptovaný formát pre LTP	Veľká Británia The National Archives	Slovenská republika Centrálny dátový archív	Česká republika Národní knihovna	Rakúsko Die Österreichische Mediathek	Nemecko Die Deutsche Nationalbibliothek	PREFORMA projekt	Dánsko Danish National Archives - Rigsarkivet	Švédsko Swedish National Archives - Riksarkivet
			všetky formáty v PRONOM						
tabuľky	sav	X	X						
	dta	X	X						
text	txt	X	X						X
	pdf, pdf/A	X	X	X		X	X		X
	html	X	X						X
	xml	X	X						X
	doc/docx	X	X						
	xls/xlsx	X	X						
geodetické dáta	mdb	X	X						
	gml		X					X	
obrázky	gif	X	X						X
	tiff	X	X			X	X	X	X
	jpeg/jpeg2000	X	X			X		X	X
	raw	X	X						
	psd	X	X						
	bmp	X	X						
	png	X	X						X
	pdf	X	X						
audio	mp3	X	X		X			X	
	aif	X	X						
	wav/wave	X	X					X	X
	ps		X			X			
video	avchd	X	X						
	mpeg-2		X		X			X	
	mpeg-4		X					X	
	dps		X						
	matroska						X		X
elektronické publikácie	epub		X	X		X			

Obr. 2 Prehľad podporovaných formátov pre LTP v niektorých európskych krajinách

Záver

Z prehľadu prístupov jednotlivých inštitúcií v Európe k formátom a LTP jednoznačne nevyplýva, ktoré formáty sú vhodné či menej vhodné pre dlhodobé uchovávanie digitálnych dát. Môžeme konštatovať, že prístup inštitúcií sa líši s ohľadom na know-how, best practice, legislatívu v danej krajine, poslanie inštitúcie, rôznorodosť projektov a pestrosť zdigitalizovaných dát. Až budúcnosť ukáže, či prístupy boli správne, či vieme spoľahlivo uchovávať dáta pre budúcnosť.

Zoznam skratiek

AIP – Archival Information Package – Archívny informačný balík
CAIR – Centrálna aplikačná infraštruktúra a registratúra
CD – Compact Disc – typ kompaktného disku
CDA – Centrálny dátový archív
IANA – The Internet Assigned Numbers Authority – register
DIP – Dissemination Information Package – Výberový informačný balík

DPX – typ formátu súboru pre video
DROID – Digital Record Object Identification
DVD – typ kompaktného disku
EPUB – typ formátu súboru pre elektronické publikácie
GIF – typ formátu súboru pre obrázky
GDFR – Global Digital Format Register
HTML – Hypertext Markup Language – hypertextový jazyk
ISO – International Organization for Standardization
JAG – typ magnetickej pásky
JPEG – typ formátu súboru pre obrázky
LTO – typ magnetickej pásky
LTP – Long Term Preservation – dlhodobé uchovávanie
MD5 – Message-Digest Algorithm je skupina kryptografických hašovacích funkcií
MP3 – typ formátu súboru pre audio
MPEG – typ formátu súboru pre video
NOC – Národné svetové centrum
OAIS – Open Archival Information System
OPIS PO2 – Operačný program informatizácia spoločnosti, Prioritná os 2 – Rozvoj pamäťových a fondových inštitúcií a obnova ich národnej infraštruktúry
PFI – pamäťová a fondová inštitúcia
PREFORMA – platforma, názov projektu
PRONOM – technický register formátov
PS – typ formátu súboru
PUID – jedinečný identifikátor formátu
RAID – Redundant Array of Independent Disks – súbor diskov – úložisko
SIP – Submission Information Package – Vkladový informačný balík
STN – slovenská technická norma
TAR – súborový formát slúžiaci na uloženie mnohých jednotlivých súborov
TIFF – typ formátu súboru pre obrázky
UC3 – California Curation Center
UDFR – Unified Digital Format Registry
UKB – Univerzitná knižnica v Bratislave
UNESCO – United Nations Educational, Scientific and Cultural Organization -Organizácia Spojených národov pre vzdelávanie, vedu a kultúru
URN:NBN – Uniform Resource Names: National Bibliography Numbers – Identifikačný systém je určený pre pamäťové a ďalšie inštitúcie na účely trvalej identifikácie digitálnych objektov v systémoch na správu, uchovávanie a sprístupňovanie dokumentov a informácií

V4 – Vyšehradská skupina alebo Vyšehradská štvorka, skrátene V4, je spoločenstvo štyroch stredoeurópskych štátov: Česka, Maďarska, Poľska a Slovenska.

ZIP – súborový formát slúžiaci na uloženie mnohých jednotlivých súborov

Zoznam odkazov

https://sk.wikipedia.org/wiki

www.iana.org

http://www.udfr.org/

http://www.nationalarchives.gov.uk/pronom/

https://polona.pl

https://academica.edu.pl

http://www.ndk.cz

https://resolver.nkp.cz

http://www.preforma-project.eu

http://cordis.europa.eu/fp7/ict/home_en.html

http://www.unesco.sk/O-UNESCO

https://www.sa.dk/en/about-us/danish-national-archives

https://riksarkivet.se

http://www.packed.be/en/

www.europeana.eu

http://www.mediathek.at/

http://www.dnb.de/DE/Home/home_node.html

http://www.nationalarchives.gov.uk

http://www.nationalarchives.gov.uk/about/commercial-opportunities/digitisation-services/our-digitisation-services/

http://www.opis.gov.sk

https://www.slovakiana.sk/

https://www.nocka.sk

http://cda.kultury.sk

Identifikace formátů – jednorázový nebo opakovaný proces?

Jan Hutař, Digital Preservation Analyst, Archives New Zealand

Anotace
Tento článek chce upozornit na potřebu opakované identifikace digitálních formátů v archivech. Identifikace formátů při vstupu dat do dlouhodobého archivu je dnes již rutinní součástí formátové strategie většiny institucí. Stejně jako všechny technologie se ale vyvíjejí i nástroje sloužící k identifikaci formátů, a mění se i globální informační infrastruktura spojená s formáty. Jednotlivé nástroje a verze nástrojů dávají odlišné výsledky a jejich přístupy k identifikaci formátů nejsou úplně srovnatelné. Důvodů pro opakovanou identifikaci může být více – nástroje a informace o formátech se zpřesňují, mění se a přibývají signatures, které k identifikaci formátů slouží, a způsob jakým se signatures používají. Na příkladu vývoje signatures v PRONOMu ukážeme, že opakování identifikace má smysl. Pokud má archiv trvale uchovat nějaký digitální obsah, měl by také sledovat změny nástrojů pro LTP, validátorů, extraktorů a především technologií používaných k identifikaci formátu souborů.

Úvod

V tomto textu se pokusím ukázat, proč by se měla opakovaná identifikace formátů stát standardní součástí formátové strategie digitálních archivů. Archivy běžně kontrolují integritu uložených souborů, měly by také opakovat identifikaci formátů. Formáty digitálních souborů jsou první věcí, nad kterou se každý, kdo vytváří a uchovává digitální data, zamyslí. Hledá se ideální formát např. pro “nejlepší obrazovou kvalitu”, “úsporu místa”, “snadné zpřístupnění”, “odolnost a použitelnost po co nejdelší dobu”. Archiváře zajímá především jak zajistit, aby obsah souborů zůstal použitelný v budoucnu, aby formát pomohl ochránit intelektuální obsah. Běžným doporučením je vybrat formát nezatížený patenty, s volně dostupnou dokumentací, podporovaný mnoha plat-

formami a aplikacemi[1]. Ne všechny formáty považované za vhodné k dlouhodobé archivaci tyto požadavky splňují. Např. specifikace formátu TIFF je od roku 1994 majetkem firmy Adobe [FLYNN, 1994].

Ne vždy lze o formátu rozhodovat jako např. v digitalizaci. Archiv, který data dostává od externích původců, může mít povinnost přijmout data, tak jak jsou. Některé archivy provádějí normalizaci do preferovaných, z hlediska dlouhodobé ochrany důvěryhodných, formátů. Jiné migrují formáty později; na základě určitých kritérií rozhodnou, kdy je nutno objekt migrovat do jiného formátu. Vždy je nutné, aby systém pro správu archivu uměl formát souboru rozpoznat.

Identifikace formátů při vstupu do digitálního archivu

Identifikace formátů, validace a extrakce technických metadat odlišuje LTP systémy (Long Term Preservation) od systémů na správu digitálních dat (DAM – Digital Asset Management). Tyto procesy jsou základem pro plánování ochranných akcí, posuzování zastarávání formátů, formátové migrace, porovnávání výsledků migrací. Identifikace formátů je pro logickou dlouhodobou ochranu souborů klíčová. Bez validace formátů či extrakce technických metadat se teoreticky lze obejít, bez identifikace formátu ale velmi těžko. Můžeme zajistit bitovou ochranu souborů, uchovat dostupnost a neměnnost souboru, to ovšem nezaručí budoucí použitelnost obsahu. Logická dlouhodobá ochrana se soustředí na ochranu intelektuálního obsahu, ten musí přežít formátové migrace. Archiv musí rozpoznat, co jsou vkládané soubory zač (formát) a získat o nich dostatečná metadata (technická, administrativní, ochranná).

Identifikace formátů je jedním z prvních kroků při vkládání souboru do LTP systému. Nejčastěji používanou aplikací pro identifikaci formátů je DROID, vyvíjený britským národním archivem (The National Archive, TNA)[2]. Dalšími jsou např. Siegfried, TriD, Apache Tika, linuxový file[3]. Nástroje určí formát souboru, ovšem výstupy i typy identifikátorů jsou různé. DROID je propojen s databází formátů PRONOM. PRONOM vznikl v roce 2004 pro vlastní potřebu TNA, a se stal nejpoužívanějším nástrojem k identifikaci formátů nedlouho poté. Identifikátory, které PRONOM (skrz DROID) přiděluje, jsou dnes základem všech LTP systémů. Tzv. PUIDy (PRONOM Unique ID), vypadají např. takto: “fmt/353” pro formát TIFF. Aplikace Siegfried také využívá

1 Viz např. doporučení Kongresové knihovny zde http://www.digitalpreservation.gov/formats/sustain/sustain.shtml

2 https://www.nationalarchives.gov.uk/information-management/manage-information/policy-process/digital-continuity/file-profiling-tool-droid/

3 http://www.itforarchivists.com/siegfried; https://tika.apache.org/; http://mark0.net/soft-trid-e.html; http://linux.die.net/man/1/file

tzv. DROID signatures (viz dále) a funkcionalitou nahrazuje DROID. Okolo DROIDu a PRONOMu vznikla komunita oborníků, kteří vytvářejí signatures formátů a posouvají nástroj neustále kupředu.

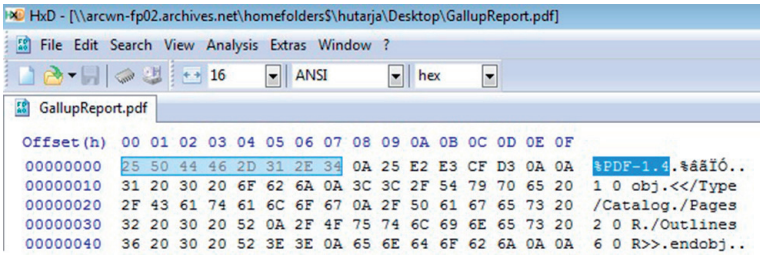
Ve většině institucí projdou data na vstupu do archivu identifikací formátů, identifikátor formátu je zapsán do metadat a proces je považován za ukončený. Archivující instituce si ale neuvědomují, že metody, nástroje a celá problematika identifikace formátů se neustále vyvíjí. Dosud se mnohem více pozornosti věnuje ingestu. Některé instituce mají celé týmy odborníků, kteří řídí vkládání dat do archivu, spravují je a provádějí ochranné akce. Pracovníci Národní knihovny Nového Zélandu (dále NLNZ) a Národního archivu Nového Zélandu (dále Archives NZ) se zamýšlejí nad opakováním procesu identifikace formátů až v posledních letech. Obě instituce začaly ukládat digitální data v LTP systému relativně "nedávno", NLNZ v roce 2008, Archives NZ v roce 2011. Od té doby se v oblasti formátů mnohé změnilo. Nezměnily se samotné formáty, ale nástroje, a hlavně již zmíněné signatures. Některé signatures byly upraveny, nahrazeny, vznikly nové. Každá nová verze XML souboru se signatures (tzv. signature file) pro DROID přináší i několik desítek nových formátů. Pokud bychom použili poslední verzi DROIDu a signature files na stejnou sbírku, která prošla identifikací formátů v roce 2008, výsledky (PUID) budou odlišné. Příkladem může být formát TIFF, jehož jednotlivé verze měly až do roku 2011 různé PUID identifikátory (fmt/7 TIFF v3, fmt/8 TIFF v4, fmt/9 TIFF v5 a fmt/10 TIFF v6). Všechny tyto PUIDy byly sloučeny do fmt/353[4]. Při opakované identifikaci mohou být některé soubory, neidentifikované v minulosti, identifikovány, pokud je přidán nový signature v databázi PRONOM.

Jak funguje DROID a proč se mění výsledky identifikace v průběhu času

DROID k identifikaci formátů používá signature file, což je XML publikované několikrát ročně na webu PRONOM[5]. Toto XML lze importovat do DROIDu. Signature file obsahuje mj. "signatures" pro jednotlivé formáty souborů, ve strojově čitelné podobě. Signature je sekvence bytů, nebo seznam více sekvencí bytů, které mohou být v konkrétním souboru určitého formátu obsaženy. Jednotlivé signature se mohou lišit konkrétností, granularitou a rozsahem. Pro některé formáty může jít pouze o sled bytů na počátku souboru, které tento formát identifikují. Např. PDF vždy začíná sledem "%PDF", v hexadecimální podobě tedy 25 50 44 46 (viz Obrázek 1), až poté následuje verze PDF – "%PDF-1.4".

4 Důvodem byly nepřesné signatures pro jednotlivé verze, jež DROID nedokázal spolehlivě rozlišit. Nový fmt/353 je obecný PUID pro všechny verze TIFF formátu.

5 https://www.nationalarchives.gov.uk/aboutapps/pronom/droid-signature-files.htm



Obrázek 1 – PDF 1.4 soubor otevřený v hexadecimálním editoru HxD

Signature může být podstatně komplikovanější, může odkazovat na řetězce bytů na počátku souboru (BOF – Beginning of the File) i na konci souboru (EOF – End of the File) a to nejen na konkrétní pozici, ale může se jednat např. o určení pozice pomocí tzv. divokých karet (*). Celý signature pro PDF verze 1.4, jak je uveden v databázi PRONOM, vypadá následovně "255044462D312E34"[6], což v hexadecimálním kódu je opravdu ono "%PDF-1.4". PRONOM uvádí, že tento řetězec je v absolutní pozici k BOF, tedy je vždy na samém počátku souboru. PUID pro PDF 1.4 je fmt/18. Příkladem formátu, jehož signature obsahuje údaje jak pro začátek (BOF) tak pro konec souboru (EOF) je JPG verze 1.02[7]. Signature také uvádí, že v každém JPG souboru je SOI (Start of the Image), který je také použit pro upřesnění. Signature každého formátu je přes interní ID propojen s popisem formátu (název, verze, PUID, koncovka), viz File Format collection v Obrázku 2.

S vydáním šesté verze DROIDu (2011) byly poprvé publikovány dva signature file. První, binární, DROID používal od počátku, druhý, nově pro kontejnerové formáty (container signature file). Odlišení binárních a kontejnerových formátů pomocí různých signature files vneslo do DROIDu nejpodstatnější změnu v jeho historii. DROID nejprve kontroluje, zda formát souboru je kontejnerový, pokud je, použije container signature file a ne binární signature file [THE NATIONAL ARCHIVES, 2011, s. 17]. Tento první krok je prováděn pomocí tzv. "trigger PUIDs", což jsou tři kontejnerové formáty, ze kterých ostatní kontejnerové formáty vycházejí. Pokud tedy souboru odpovídá signature jednoho z těchto tří PUIDů, DROID usoudí, že formát je kontejnerový

6 http://www.nationalarchives.gov.uk/PRONOM/Format/proFormatSearch.aspx?status=detailReport&id=617&strPageToDisplay=signatures

7 http://www.nationalarchives.gov.uk/PRONOM/Format/proFormatSearch.aspx?status=detailReport&id=669&strPageToDisplay=signatures

a pokusí se jej identifikovat přesněji[8]. Kontejnerové signatures jsou priorizovány před běžnými signatures. Trigger PUIDy, OLE2 formát (fmt/111) a dva ZIP formáty (fmt/189 a x-fmt/263), jsou na konci kontejnerového signature file[9] (viz Obrázek 2). Kontejnerové signatures jsou navrženy tak, aby byly přesnější než binární. Zavedení kontejnerových signatures přineslo i komplikace, kontejnerové signatures nejsou dostupné ve veřejné databázi PRONOM[10]. Datový model PRONOMu se nezměnil a kontejnerové signatures obsahují pole, která v databázi PRONOM nejsou, jsou proto dostupné jen v XML souboru. Kontejnerový signature file je závislý na binárním signature file, kontejnerové formáty obsahují link na jejich popis obsažený v binárním signature file. Kontejnerové formáty zanesené do PRONOMu před vznikem kontejnerového signature file mají v binárním signature file i starou verzi signature (příklad je MS WORD fmt/40, internal signature ID 182), nová je v kontejner signature file. Nové kontejnerové formáty mají v binárním signature file pouze popis, signature je jen v kontejnerovém signature file (např. fmt/677, Serif PagePlus).

Databáze PRONOM obsahuje (srpen 2016) celkem 1403 záznamů formátů. Ne všechny formáty v PRONOMu mají signature, některé nemají a jsou jen prázdnou schránkou. Mají ale PUID, který lze použít, pokud chceme tyto formáty popsat, případně použít k identifikaci na základě přípony souboru. 924 formátů má signature, 429 signature nemá. Stav se průběžně mění s každým vydáním signature file. Uvedená čísla jsou platná pro signature file verze 86 z července 2016. Ten přinesl 46 zcela nových formátů, 23 formátů bylo aktualizováno a přidáno bylo 46 signatures[11]. Nové formáty a signatures nejsou nutně totožné, číslo 46 v tomto vydání signature files se shoduje, ale při bližším zkoumání je jasné, že PUIDy se neshodují. Tj. byly přidány nové formáty bez signature a některé formáty již existující v PRONOM databázi dostaly vlastní signature. Tento postup je běžný.

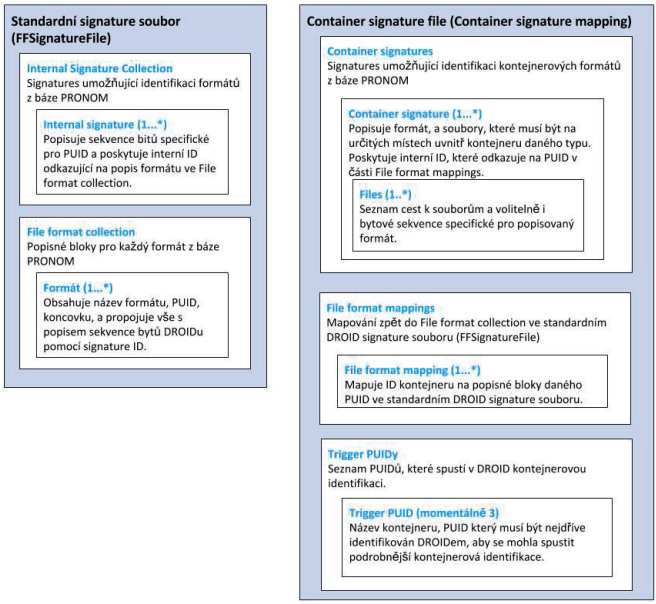
Identifikace formátů nemůže být jednorázovou aktivitou. Mění se prostředí, nástroje, ale především signature a záznamy formátů, vznikají nové signature, jsou zpřesňovány, deaktivovány, nahrazovány.

8 Např. MS Office formáty založené na xml, tedy docx, pptx a xlsx jsou ZIP kontejnery.

9 Viz aktuální container signature file dostupný zde https://www.nationalarchives.gov.uk/documents/container-signature-20160727.xml

10 https://www.nationalarchives.gov.uk/PRONOM/BasicSearch/proBasicSearch.aspx?status=new

11 Podrobnosti o novinkách v každém signature file jsou http://www.nationalarchives.gov.uk/aboutapps/pronom/release-notes.xml



Obrázek 2 – struktura binárního (vlevo) a kontejnerového signature file pro DROID [SPENCER, 2016]

Jak může opakovaná identifikace formátů probíhat?

Aktivní dlouhodobá ochrana spočívá mj. ve vytváření metadat v průběhu životního cyklu souboru. Nutnost udržovat metadata v aktuální podobě platí i pro identifikaci formátů. Identifikátor formátu získaný během vkládání do digitálního archivu nemusí v budoucnu platit. Opakovanou identifikaci formátů lze provádět různě. Například můžeme jednou za 5 let provést identifikaci formátů všech souborů v archivu. Pro menší archivy to může být schůdné, většinou ale narazíme na problém s množstvím souborů, nároky na výpočetní výkon a čas. Archives NZ má v LTP systému asi 4,5 milionů souborů. Identifikace všech souborů by byl menší projekt vyžadující zdroje, plánování, monitorování apod. Opakovaná identifikace formátů nesmí brzdit běžné procesy v LTP systému. Druhou možností je provádět opakovanou identifikaci formátů výběrově na části obsahu digitálního archivu vybrané na základě nějakých kritérií (obrazové soubory, data z konkrétního projektu, od určitého producenta, soubory uložené před určitou dobou). Nebo lze na základě analýzy archivu identifikovat soubory s PUIDy, které byly v posledním vydání signature files nahrazeny, upraveny či zrušeny a tyto podrobit nové identifikaci formátů. Tento inkrementální přístup nezahrne formá-

ty, které jsou v PRONOMu nové, případně formáty, které neměly signature a nově jej mají. Teoreticky by to šlo ošetřit, ale prakticky je to řešitelné pouze v případě, že opakovaná identifikace pomocí nových signature files je provedena pro kompletní obsah digitálního archivu. Má-li instituce objekty, jejichž formát je "neznámý" nebo „nejistý", je velká šance, že opakovaným procesem identifikace se tato množina zmenší.

Co opakovaná identifikace formátů znamená pro systém digitálního archivu
Různé systémy mají odlišné možnosti jak opakovanou identifikaci formátu provést. LTP systémy jako Rosetta či Preservica jsou na tento typ aktivit připraveny. Je v nich možné na základě parametrů vytvořit množiny dat, nastavit sled kroků (proces), a proces pak aplikovat na vytvořený set. Jedním z kroků může být identifikace formátu. Pak jsou systémy, jako např. Archivematica, která v současné verzi výše popsané neumožní a jedinou možností je re-ingest dat.

Výsledek opakované identifikace formátů by se měl promítnout do metadat. LTP systém Rosetta při jakékoliv aktivitě provedené na AIP balíčku vytvoří metadata o provedené události. Při opakované identifikaci formátu dojde k připsání údajů o formátu do metadat, jsou aktualizovány informace o použitém nástroji a jeho verzi (DROID), a použité verzi signature file. Původní PUID je v metadatech zachován, stejně jako informace o původní identifikaci formátu. Do metadat je přidána událost nové identifikace formátů, kdy proběhla, kdo ji spustil a jaký byl výsledek. Takto zůstává zachována informace o všech výsledcích identifikace formátu konkrétního souboru. Opakování procesu identifikace formátů je v LTP systému Rosetta důvodem pro vytvoření nové verze XML s metadaty AIP balíčku. Soubory se neduplikují, původní XML s metadaty AIP je také uchováno. Jiné systémy toto mohou řešit různě.

Je opakovaný proces identifikace formátů v silách všech institucí?
Ne všechny instituce budou mít kapacitu se do opakované identifikace formátů pustit. Systém, který používají, nemusí být z různých důvodů na něco takového připraven. Hlavní překážky ale budou personální. Na mnoha datech je opakovaná identifikace formátů proveditelná spíše formou projektu, a k tomu je třeba vyčlenit zaměstnance. Menší instituce často dělají pouze to nejnutnější, řeší ingest, správu archivu a dat, na nic dalšího zdroje nemají. Stávající zaměstnanci navíc nemusí mít potřebné znalosti (více násobná identifikace, neznámé formáty). Užitečné mohou být komunity uživatelů DAM nebo LTP systémů. Sledují vývoj okolo formátů a prosazují změny v konkrétních systémech např. pomocí formátových knihoven (Archivematica, Rosetta, Preservica). Záležet bude také na povaze sbírky. Pro homogenní obsah, tj. např. z interní digitalizace nebo od několika málo původců, kteří mají jasně dané standardy pro tvorbu dat, může být plánování a provedení opakované identifikace podstatně jednodušší.

Archives NZ a National Library of New Zealand

Obě organizace se věnují dlouhodobé ochraně digitálních dat a mají dedikované týmy. NLNZ používá LTP systém Rosetta od roku 2008, Archives NZ od roku 2011. NLNZ a Archives NZ sdílí jednu instanci systému Rosetta, včetně většiny infrastruktury. Každá instituce je spravována nezávisle, ingest workflow jsou odlišná, nastavení je plně pod kontrolou institucí. O opakování procesu identifikace obě instituce uvažovaly již několikrát, vždy se ale našly jiné věci, které dostaly přednost. Nyní, 8 a 5 let po začátku používání systému, ale není možné novou identifikaci dále odkládat. Testování a plánování probíhá od roku 2015, realizace je plánována na rok 2017 s tím, že obě instituce budou identifikaci provádět nezávisle. NLNZ má 9 milionů souborů, které dohromady tvoří asi 1,5 milionu intelektuálních entit. Archives NZ má 4,5 milionu souborů, asi 220 tisíc entit. Velikost obou sbírek je okolo 120 TB na každé straně. Hlavním důvodem proč se celý proces bude odehrávat nezávisle a jinou metodou v každé instituci je struktura dat. NLNZ má podstatně více formátů, 162 různých PUID. Pro srovnání v roce 2013 to bylo 123 formátů [SAJWAN, 2013, s. 3]. Archives NZ má v tuto chvíli 39 různých formátů. NLNZ vedle digitalizace akceptuje data od externích vydavatelů v jakémkoliv formátu. V Archives NZ většina dat pochází z interní digitalizace, jde o omezený počet formátů. Transfery born digital archiválií ze státní správy proběhly zatím pouze několikrát, proto není počet formátů tak vysoký. Pravidelné born digital transfery začnou v roce 2017. Ani Archives NZ nelimituje původce specifikací několika málo vstupních formátů. NLNZ začala Rosetta používat v roce 2008 a z tohoto období pochází množina souborů, které mají neznámý formát, případně byla učiněna rozhodnutí, která by dnes nebyla akceptovatelná. NLNZ se proto rozhodla opakovat identifikaci formátů pro všechny soubory v jejich archivu. Archives NZ půjde cestou výběrového opakování, pro určité formáty a sbírky. 56% souborů v Rosettě Archives NZ je JPEG 2000, 40% jsou TIFF soubory, vše z interní digitalizace. Již nyní víme, že není potřeba provádět opakovanou identifikaci pro JPEG 2000 soubory, výsledek by byl stejný jako je současný PUID. Pro soubory TIFF je situace jiná, spousta z nich má PUID fmt/7, který byl v roce 2011 nahrazen za fmt/353, pro ně je nová identifikace žádoucí. Další soubory pocházejí z digitálních transferů, a byly ingestovány v roce 2015 a 2016, opakování identifikace v tuto chvíli není potřeba.

Po ukončení opakované identifikace v obou institucích bude následovat rozhodnutí, jak dál. Nelze čekat několik let a pak opět provést identifikaci pro celý obsah LTP systému, nemusí to být proveditelné kvůli množství dat. Počítá se s tím, že každé vydání DROID signature file bude analyzováno tak, abychom zjistili, jaké změny pro konkrétní formáty byly provedeny a na základě této analýzy budeme provádět opakovanou identifikaci formátů na souborech s relevantním PUID (nebo PUIDy). Do jisté míry se bude jednat o neustálý inkrementální proces identifikace formátů.

Závěr

Identifikace formátů je klíčový proces dlouhodobé ochrany digitálních dat. Tento proces by se měl v digitálním archivu opakovat, stejně jako např. kontrola fixity. Celý proces opakované identifikace formátů přináší spoustu otázek, bezpochyby je personálně, časově i výpočetně náročný. I přesto by se měl stát běžnou součástí životního cyklu archivovaných souborů.

Použitá literatura

FLYNN, Laurie. Aldus and Adobe Lay Claim to Digital Publishing. *The New York Times*. 24. Srpen 1994 [cit. 2016-09-27]. Dostupné také z: http://www.nytimes.com/1994/08/24/business/business-technology-aldus-and-adobe-lay-claim-to-digital-publishing.html

The National Archives. *DROID: How to use it and how to interpret your results* [online]. The National Archives, 2011 [cit. 2016-09-27]. 36 s. https://www.nationalarchives.gov.uk/documents/information-management/droid-how-to-use-it-and-interpret-results.pdf

SAJWAN, Suman. *Securing the future: Digital Preservation at the National Library of New Zealand* [online]. National Library of New Zealand: Wellington, 2013 [cit. 2016-09-27]. 8 s. https://digitalpreservation.natlib.govt.nz/assets/NDHA/Publications/2013/NDHA-Booklet-FINAL.PDF

SPENCER, Ross. DROID Container Signature Files: What they are and how to create them: A template and an example, or few… . In *Open Preservation Foundation* [online], 7.1. 2016 [cit. 2016-09-27]. http://openpreservation.org/blog/2016/01/07/droid-container-signature-files-what-they-are-and-how-to-create-them-a-template-and-an-example-or-few/

Formátová strategie LTP úložiště NK ČR

Ladislav Cubr, Národní knihovna ČR

Abstrakt
Formátová strategie není jen výběr vhodného formátu, ale řada dalších souvisejících aktivit, mezi nimiž hraje klíčovou roli validace.

Problematiku souborových formátů s ohledem na dlouhodobé uchovávání (LTP) začala Národní knihovna ČR soustavněji řešit až v souvislosti s budováním LTP úložiště v projektu Vytvoření Národní digitální knihovny.
Příspěvek představuje současnou formátovou strategii LTP úložiště NK ČR.

1 Úvod

Národní knihovna ČR (NK ČR) byla v letech 2009-2014 příjemcem projektu Vytvoření Národní digitální knihovny (projekt NDK), který byl financován z Integrovaného operačního programu EU programového období 2007-2014. NK ČR se v souvislosti s tímto projektem začala poprvé soustavněji věnovat digitální archivaci (dlouhodobému uchovávání digitálních dokumentů) podle normy ISO 14721. Ta předpokládá provoz digitálního repozitáře, tj. systému, který archivaci podporuje (1). Takovýto repozitář (LTP úložiště) byl jedním z výstupů projektu NDK. V listopadu 2011 byla vydána nová zřizovací listina NK ČR, v jejímž článku II.[1] je již explicitně uveden závazek digitální archivace i odkaz na koncept důvěryhodného digitálního repozitáře (2).[2]

V rámci NK ČR má odbornou a kurátorskou stránku digitální archivace na starosti specializovaný odbor (ODIF, Odbor digitálních fondů).[3] ODIF při přípravě projektu

1 „Formuluje strategie a postupy dlouhodobé ochrany elektronických dokumentů a provozuje důvěryhodné digitální úložiště."

2 V té době citovaná norma ISO 16363 ještě nevyšla, nicméně byla v přípravě jako přepracovávaný de facto standard TRAC (23).

3 Na počátku nesl název Odbor digitální ochrany.

NDK zavedl nové standardy a postupy ve třech klíčových oblastech digitální archivace (metadata, trvalé identifikátory a datové formáty). Pro metadata vytvořil standard NDK (metadatový aplikační profil pro digitalizaci).[4] Profil je určen k zaznamenávání metadat v průběhu digitalizace tištěných dokumentů, která jsou důležitá z hlediska archivace. Pro trvalou identifikaci digitálních dokumentů odbor navrhl a vybudoval systém nazvaný ČIDLO, který je založen na standardu URN:NBN (3, s. 41-43). Jeho součástí je i národní resolver.[5] Datovým formátům je pak věnován tento článek.

2 Datové formáty a digitální archivace

Hlavním cílem dlouhodobého uchovávání (digitální archivace) je podle normy ISO 14721 zachování informačního obsahu (content information), který je reprezentován datovým objektem s obsahem (content data object) (1, s. 27). Za účelem zachování informačního obsahu v dlouhodobém horizontu je nutné průběžně získávat a mít stále k dispozici dostatečné interpretační informace (representation information), tj. informace nezbytné ke správné interpretaci datového objektu s obsahem. V digitálním světě mezi datový objekt s obsahem a lidského uživatele nevyhnutelně vstupuje také technologie, jež zahrnuje software i datové formáty. Datový formát je jedním z typů interpretačních informací, v tomto případě o tom, jak data v daném formátu reprodukovat. V ideálním případě jednak existuje dostatek volně dostupných (specializovaných) softwarových aplikací, které znalostí daného formátu disponují a dokáží s ním pracovat, jednak jsou informace o formátu obsaženy v dostupné a dostatečně dobře popsané dokumentaci (formátové specifikaci). Počítačová realita má však k tomuto ideálu často daleko a právě datové formáty jsou jednou z oblastí, ve kterých se skrývají největší rizika pro digitální archivaci. V řadě případů není k formátům dostupná specifikace nebo se k nim váží licenční omezení, což představuje rizika pro uchovávání dat. Ve všech případech pak představuje (pro laiky paradoxně) největší problém rychlý vývoj informačně komunikačních technologií, který s sebou nese hrozbu zastarávání datových formátů. Jeden z odhadů průměrné délky zastarání formátu od doby uvedení na trh je 8-20 let (4). Zastarávání se projevuje jednak rizikem ztráty dostupnosti softwarových aplikací, které dokáží formáty adekvátně reprodukovat (tj. zobrazit, přehrát nebo jiným způsobem prezentovat smyslům lidského uživatele), jednak ztrátou schopnosti nové generace softwaru formát zpracovávat (převádět do jiného formátu apod.).

4 http://www.ndk.cz/standardy-digitalizace/metadata

5 https://resolver.nkp.cz/

Formátovou strategii lze pojímat jako soubor pravidel a postupů pro řízení práce s datovými formáty v celém životním cyklu datových objektů s obsahem a jako takovou ji lze zařadit k jedné z hlavních složek celkové strategie digitální archivace. Soubory pravidel by měly vymezovat minimálně následující okruhy: výběr formátu, dohodu o dodávání dat, formátové konverze (včetně formátové migrace jako zvažovaného budoucího opatření digitální archivace), metadata a kontrolu integrity formátu.

Základním principem strategie LTP úložiště NK ČR je řídit se doporučenými postupy mezinárodní knihovnické komunity a inspirovat se osvědčenou praxí významných zahraničních paměťových institucí, případně mezinárodní standardy pro naše prostředí lokalizovat (v mezích, které tyto standardy umožňují).

3 Výběr datového formátu

Pro formátovou strategii je prvním a zásadním krokem výběr vhodného datového formátu. Z hlediska způsobu užití je při správě digitálních dokumentů třeba odlišovat dvě skupiny formátů – archivační formáty (určené pro dlouhodobé uchovávání) a prezentační formáty (určené pro zpřístupňování čtenářům a dalším uživatelům). Prezentační formát je volen s ohledem na komfort čtenářů a současné možnosti přístupu k informačnímu obsahu. Volba tedy závisí především na podpoře v internetových prohlížečích a rozšířenosti formátů mezi uživateli, případně dalších ukazatelích, které vycházejí z požadavků a omezení běžného užití v aktuálně široce dostupném počítačovém prostředí. Archivační formát je volen z hlediska jeho (aktuální) vhodnosti pro uložení datových objektů s obsahem v balíčku AIP v digitálním repozitáři za účelem dlouhodobého uchovávání. Cílem je uložit obsah v takovém formátu, o kterém se předpokládá, že jeho užití v současnosti a blízké budoucnosti nebude představovat větší riziko. V případě, že vkladatel do repozitáře dodá data v jiném formátu než archivačním, je doporučeným postupem, aby repozitář provedl normalizaci do archivačního formátu. Pro volbu archivačního formátu existují doporučení uznávaných organizací. Mezi kritéria, která se objevují nejčastěji, patří vesměs otevřenost formátu (tj. dostupnost formátové specifikace) a nezatíženost patenty [viz např. (5), (6)]. V době přípravy projektu NDK bylo nejcitovanějším doporučení Floridského digitálního archivu (Florida Digital Archive) obsahující seznam nejběžnějších formátů s hodnocením jejich spolehlivosti z hlediska archivace (na třístupňové škále – vysoká, střední, nízká).[6] Kongresová knihovna (Library of Congress) začala jednou za rok vydávat seznam doporuče-

6 Nazvaný „Recommended Data Formats for Preservation Purposes in the Florida Digital Archive“. V této původní podobě již není dostupný [citováno podle (11, s. 84)].

ných formátů (jak pro digitální, tak pro fyzické objekty). Poslední seznam byl vydán letos v červenci (7). Z roku 2014 pochází srovnání vybraných obrazových formátů americkou iniciativou FADGI[7], a to na základě poměrně velkého počtu kritérií seskupených do čtyř hlavních kategorií (udržitelnost; ekonomické faktory; požadavky na implementaci; nastavení a možnosti) (6).

V souvislosti s přijetím modelu OAIS jakožto východiska pro digitální archivaci se v komunitě paměťových institucí objevil navazující koncept registrů interpretačních informací (8) (v širší knihovnické komunitě jsou označovány za formátové registry). Ty by podle navrhovatelů měly zaznamenávat to, co model OAIS označuje jako síť interpretačních informací (representation network) (1, s. 51-52), v naší praxi tedy zejména informace o datových formátech, souvisejících aplikacích a všech ostatních prvcích počítačového prostředí, které podporuje adekvátní reprodukci digitálních objektů a jejich zpracování. Za tímto návrhem stála pragmatická úvaha, podle níž nejsou jednotlivé instituce schopny všechny potřebné interpretační informace dokumentovat vlastními silami. Nejstarším registrem je PRONOM, který je provozován britskými Národními archivy (The National Archives) téměř patnáct let.[8] Registr obsahuje poměrně velkou databázi datových formátů, jejich záznamy jsou však často minimální. Ačkoliv PRONOM zdaleka nesplňuje své původní ambice, jde v současnosti de facto o jediný důvěryhodný projekt, který se alespoň snaží uskutečňovat původní vizi směrodatného globálního registru interpretačních informací. Především však (jako jediný registr vůbec) nabízí jednoznačný a jedinečný identifikátor datového formátu (přesněji řečeno jednotky interpretačních informací, z nichž za nejdůležitější je nyní považován právě datový formát), identifikátor PUID (9). Ten zohledňuje i odlišné verze nebo různé podtypy formátů.[9] Takto členěná identifikace je z hlediska digitální archivace klíčová. Různé verze formátu mohou být svázány s odlišnými riziky. Za druhý významný zdroj lze považovat také registr Kongresové knihovny, který je rovněž udržován.[10] Ačkoliv obsahuje méně záznamů, jsou podrobnější. Dva registry podobné úrovně již žel zanikly.[11] To ukazuje na křehkost takto koncipovaných projektů, jejichž význam je nezastupitelný – typický jev v komunitě paměťových institucí.

Hlavním typem digitálních dat, se kterými NK ČR již dlouhodobě pracuje, jsou rastro-

7 Federal Agencies Digitization Guidelines Initiative

8 http://www.nationalarchives.gov.uk/PRONOM/

9 Například PUID pro JPEG verze 1.00 je „fmt/42“, pro verzi 1.01 „fmt/43“ a pro verzi 1.02 „fmt/44“. Dosud stále nejužívanější obecný registr MIME odlišuje formáty jen na základě názvu.

10 http://www.digitalpreservation.gov/formats/

11 GDFR a UDFR.

vá obrazová data (tvořená především digitalizací tištěných dokumentů). Jako archivační formát byl před projektem NDK užíván JPEG a jako prezentační formát DjVu. Při přípravě projektu bylo třeba vhodný formát znovu zvážit, s ohledem na výši investice, masový záběr plánované digitalizace a vývoj v oblasti digitální archivace. Průzkumem velkých zahraničních digitalizačních projektů bylo zjištěno, že nejužívanějším archivačním formátem je TIFF a prezentačním JPEG (10, s. 64). V té době se v zahraničních paměťových institucích začalo rozšiřovat využití formátu JP2 (tehdy relativně nového formátu pro rastrové obrazy), a to jak pro archivaci, tak zpřístupnění. Zmíněný seznam Floridského digitálního archivu uváděl jako rastrová obrazová data s vysokým stupněm spolehlivosti formáty JP2 (bezeztrátová komprese), TIFF (bez komprese) a PNG, zatímco „klasický" JPEG a JP2 (ztrátová komprese) zařadil do kategorie se středním stupněm spolehlivosti [cit. dle (11, s. 84)]. V roce 2008 byla také publikována vlivná studie srovnávající formáty JP2, TIFF a JPEG z hlediska potřeb dlouhodobého uchovávání – jejím vítězem se stal právě JP2 (12). Již tehdy bylo také zřejmé, že formát DjVu jako prezentační formát (jako archivační nebyl zvažován nikdy) již zastaral.[12] Výše uvedené skutečnosti měly vliv na výběr formátu JP2 jako nového archivačního i prezentačního formátu pro NDK, přičemž za jeho hlavní výhody byly označeny otevřená dokumentace, neproprietárnost, kompresní možnosti a využitelnost pro archivaci i zpřístupnění (10, s. 64). Současně je však nutné poznamenat, že implementace formátu JP2 pro produkci a zpřístupnění byla a stále je vysoce náročný proces (6). Přinejmenším ve fázi implementace je důležité zapojení specialisty. Řada zahraničních knihoven si za tímto účelem najala Roberta Buckleyho (spoluautora specifikace formátu), například Kongresová knihovna pro americký program digitalizace novin (13). Jedním z úkolů specialisty je zvolit vhodný profil (specifikaci požadovaných vlastností formátu), přičemž parametrizace formátu JP2 není triviální operace. Některé zahraniční instituce si zvolily pouze jeden profil (tj. tentýž pro archivaci i prezentaci) (14). V projektu NDK byl zvolen přístup s různými profily. Úspěchem byla možnost využít matematicky bezeztrátovou kompresi pro archivační formát. Celá řada institucí, která v současnosti přechází nebo se chystá přecházet na JP2 jako archivační formát (konverzí z formátu TIFF), totiž volí matematicky ztrátovou kompresi, a to z důvodu významné úspory úložných kapacit [např. (14), (15)]. I když jde o vizuálně bezeztrátovou kompresi (tedy uživatel nic nepozná), ztrátově komprimovaný formát je vždy rizikem pro budoucí migrace.

Jistou nevýhodou JP2 jako prezentačního formátu je skutečnost, že není nativně podporován v prohlížečích, takže je do prohlížeče nutné nainstalovat plugin. Je však možné využít implementační variantu s obrázkovým serverem, která byla zavedena i v NDK. Profil JP2 pro zpřístupnění slouží jako produkční matrice (production mas-

12 Jedním z důvodů byla prohra souboje s jeho hlavním konkurentem, formátem PDF.

ter) pro vytváření dočasných obrázků ve formátu JPEG pro uživatele při jeho procházení digitální knihovnou (10). Tím je vyřešen problém s nutností pluginu.

V rámci plánovaného přijímání elektronických publikací vydavatelů k dlouhodobému uchovávání v LTP úložišti byla v rámci projektu „Správa elektronických publikací v síti knihoven České republiky“[13] vytvořena doporučení na požadované archivační formáty. Na základě průzkumu zahraniční praxe byly zvoleny archivační formáty PDF/A-1, PDF/A-2 a ePub2 (16, s. 19-22). V případě, že vydavatelé budou dodávat do NK ČR publikace v těchto formátech, může jim LTP úložiště garantovat jejich dlouhodobé uchovávání.

LTP úložiště se rovněž připravuje na možnost uchovávání digitálních zvukových dokumentů. V rámci výzkumu NK ČR vznikla analýza zahraniční praxe. Podle ní se zatím vhodnými kandidáty jeví být formát WAV jako archivační formát a mp3 jako prezentační. Průzkum ukázal, že jde o nejčastěji užívané formáty a že existuje velké množství dostupných nástrojů pro práci s nimi (17). Na základě této analýzy bude rozhodnuto o požadovaných archivačních formátech pro zvukový obsah.

4 Dohoda o dodávání dat

Podoba dat, která vkladatel dodává do repozitáře za účelem dlouhodobého uchovávání, je podle normy ISO 14721 věcí dohody o dodávání dat (submission agreement) mezi repozitářem („archivem OAIS“) a vkladatelem dat (často je jím sám producent dat) (1). To zahrnuje i požadavky na metadata, identifikátory a především datové formáty. Na základě takové dohody může vkladatel svá data sám převést do archivačního formátu a v této podobě je dodat do repozitáře. Tím repozitáři ušetří čas i nemalé prostředky. Druhou možností je, že repozitář provede při příjmu dat formátovou normalizaci, tj. převod do preferovaných archivačních formátů. Je realitou, že v praxi v řadě případů může být taková dohoda velmi obtížná (viz případ nedostatečných legislativních ustanovení). Repozitář však v každém případě musí při fázi příjmu dat věnovat zvláštní pozornost tomu, v jakých formátech jsou data ukládána do balíčků AIP.

NK ČR do svého LTP úložiště začala přijímat také digitalizované tištěné dokumenty příjemců podpory z podprogramu VISK7. Tato situace je pro NK ČR výhodná v tom smyslu, že vzhledem ke svému postavení metodického centra a podmínkám VISK7 může jiným knihovnám předepisovat specifikaci datových formátů. Využití

13 http://edeposit.nkp.cz/

možnosti určit za formát dodávaných dat právě formát archivační je jedním ze základních pilířů formátové strategie našeho LTP úložiště. Pro NK ČR tím odpadá nutnost provádět (časově, technicky i finančně) náročnou formátovou normalizaci. Na druhou stranu to znamená větší nároky na producenty dat. Jednotná pravidla však přináší i výhody a úspory díky jednotnosti systému a pravidel v celém úseku. Součástí strategie LTP úložiště je i aktivní podíl na návrhu podmínek podprogramu VISK7, specifikujících dodávaná data (mj. profily pro JP2) a způsob jejich dodávání. Podmínky podprogramu VISK7 také v současnosti nahrazují dohodu o dodávání dat. ODIF však připravuje návrh samostatně uzavírané dohody o dodávání dat (tedy nad rámec projektových smluv VISK7).

5 Formátové konverze

Velkou pozornost je třeba věnovat volbě vhodných nástrojů pro konverzi do požadovaného archivačního formátu. Výjimkou jsou případy, kdy požadovaný archivační formát je totožný s prvním formátem životního cyklu. Častým příkladem je skenování do formátu TIFF, který pak zůstává i archivačním formátem [viz např. (18)]. Je záhodno, aby repozitář doporučil vkladateli vhodné nástroje, zejména pokud jich existuje více, a v ideálním případě pak takové, které jeho pracovníci sami otestovali. Bez ohledu na sebelépe zvolený archivační formát bude každý repozitář jednou muset řešit případ zastarání formátu, ve kterém byl datový objekt s obsahem uložen v první verzi balíčku AIP. V případě, že si repozitář jako součást své formátové strategie zvolil formátovou migraci, je otázka aktuálně vhodných konverzních nástrojů předmětem sledování (stejně jako jím je zastarávání formátu) a následného testování.

V projektu NDK i následných digitalizačních projektech, které se řídí standardy LTP úložiště, jsou skenované stránky vytvářeny ve formátu TIFF (bez komprese). Pro konverzi formátu TIFF do JP2 existuje několik nástrojů (kodeků). Široce užívaným ve velkých zahraničních institucích je komerční produkt firmy Kakadu (19). K dispozici jsou již také volně dostupné opensourcové nástroje, mezi nimiž má kodek OpenJpeg asi nejširší uživatelskou základnu a je považován za největšího konkurenta Kakadu. NK ČR na základě testů provedených v minulosti používá (pro svoji produkci) a doporučuje (pro digitalizaci jiných knihoven) kodek Kakadu. Na webu NDK je také zveřejněn příkazový řádek pro konverzi do Kakadu.[14] Knihovny však podle podmínek podprogramu VISK 7 mohou využít i jiné nástroje. Část knihoven používá kodek Kakadu, část OpenJpeg. V minulosti testovaná verze kodeku OpenJpeg nebyla v ODIF shledá-

14 http://www.ndk.cz/standardy-digitalizace/standardy-pro-obrazova-data

na za dostatečně kvalitní. Pozdější verze kodeku však přinesly zlepšení a například ve srovnání provedeném v roce 2013 specialisty z British Library se OpenJpeg umístil před Kakadu (15). Práce na tomto otevřeném kodeku pokračují a stále jsou vydávány verze. NK ČR se proto chystá provést podrobný test nejnovější verze OpenJpegu, případně pak na webu NDK zveřejnit příslušné příkazové řádky.

V projektu Správa elektronických publikací v síti knihoven České republiky byl proveden průzkum mezi českými vydavateli. Zjistilo se, že hodně z nich využívá pro produkci elektronických publikací nástroje Adobe InDesign, Microsoft Word a Open Office. Pro tyto nástroje byla tedy vytvořena doporučení pro nastavení převodu do archivačního formátu PDF/A (16, s. 23-34). Rovněž byl doporučen nástroj pro přípravu publikací ve formátu ePub2.

V průzkumu zvukových dokumentů bylo zjištěno, že pro konverzi z formátu WAV do formátu MP3 existuje poměrně velké množství nástrojů, a to jak komerčních, tak volně dostupných (17). Konverzi je možné provést různými nahrávacími a editačními nástroji, případně samostatnými konvertory. Nyní se připravuje testování několika volně dostupných verzí a demoverzí placených nástrojů. Přinejmenším budou otestovány nahrávací a editační nástroje Sound Forge, Audacity, WaveLab, Adobe Audition a konvertory LameXP a FFmpeg. Na základě výsledků testů a konzultací se zvukovými odborníky pak budou vybrány vhodné nástroje.

NK ČR má ve svých fondech také velký objem digitalizovaných dokumentů vzniklých před projektem NDK. Vzhledem k tomu, že pro rastrová obrazová data je v naší strategii zatím definován jediný archivační formát (JP2), je cílem LTP úložiště provést jejich normalizaci. V současnosti probíhá projekt na převod těchto dat do JP2, pro který je využívána infrastruktura současné digitalizace. Rozdíl je v tom, že na počátku postupu je vsunut krok zahrnující převod z formátu JPEG do TIFF. Současně platí, že pro celý proces musejí být data kompletována spolu s metadaty uloženými ve starém systému pro zpřístupnění,[15] což se neobjede bez značného objemu manuální kontroly. V některých případech se zjistilo, že došlo ke ztrátě archivních kopií. Zde nezbývá než generovat soubory JP2 z prezentačního formátu (tedy v posloupnosti DjVu–TIFF–JP2). Celý projekt je de facto prvním případem formátové migrace (normalizace) jakožto ochranného opatření, které provádí naše LTP úložiště.

15 http://kramerius.nkp.cz/

6 Metadata

Z hlediska zachování autenticity je nezbytné zaznamenat celou historii datového objektu s obsahem, včetně jeho původu, a to jak staticky (popis jednotlivých generací těchto objektů, od původních souborů ve formátu TIFF až po konečný JP2 v archivačním profilu), tak dynamicky (popis událostí provedených na těchto objektech, včetně činitelů těchto událostí). Je tedy nutné zaznamenat mj. technické informace o jednotlivých datových generacích (např. původní skeny v TIFF se mažou a zůstává po nich jen právě ona „stopa" v metadatech), o provedených operacích (např. o konverzi z TIFF do JPEG) a jejich činitelích (např. o užitém kodeku nebo snímacích zařízeních).

Pro digitalizaci NDK byly vytvořeny dva metadatové aplikační profily vycházející z mezinárodních metadatových standardů, založené na kontejnerovém formátu METS. Koncept aplikačního profilu je založen na myšlence, že pro konkrétní potřebu je nutno metadatové standardy lokalizovat a optimalizovat (20, s. 54). Jeden profil byl vytvořen pro monografie, druhý pro periodika. Od začátku projektu NDK do současnosti bylo vydáno několik verzí profilů, pro periodika je v současnosti poslední verzí 1.6, pro monografie 1.2. Všechny verze i historie změn jsou dokumentovány a zpřístupňovány na webu NDK.[16] Pro záznam výše uvedených informací o historii datového objektu s obsahem (zde tedy konkrétně rastrových dat) jsou klíčové standardy PREMIS (pro popis událostí, činitelů a částečně i objektů) a MIX (specificky pro popis rastrových objektů). Vybrané položky obou standardů tvoří součást metadatového profilu NDK. Technické informace o rastrových datech jsou získávány v NK ČR dvěma nástroji – DROID (přidělení identifikátoru PUID) a JHOVE (další technické informace), přičemž jejich výstupy jsou následně zaznamenány do metadat.

Aktuálně se také pracuje na metadatovém aplikačním profilu pro zvukové dokumenty. Bude využit kontejner METS a standard PREMIS podobně jako u rastrových obrazů, pro podrobnější technická metadata bude užit metadatový standard AES57 (17). Nejasnosti zatím panují ohledně zápisu metadat týkajících se vzniku zvukového dokumentu. Práce na standardu, který měl tento zápis umožnit, se totiž zdají být zastaveny.[17] Hledá se tedy nový způsob, jak tyto údaje (standardizovaně) zaznamenat.

V projektu Správa elektronických publikací v síti knihoven České republiky byl vytvořen metadatový aplikační profil pro sběr elektronických dokumentů (16, s. 58-60). Metadata nebudou vytvářena vkladateli (vydavateli), ale v našem LTP úložišti (speciální aplikací, která byla navržena pro vytváření balíčku AIP z publikací dodávaných

16 http://www.ndk.cz/standardy-digitalizace/metadata

17 http://www.aes.org/standards/meetings/project-status.cfm

vydavateli). Pro popis formátu je užit PREMIS a pro zápis některých specifických informací o formátech PDF a EPUB bylo navrženo vlastní schéma LTP úložiště.

7 Kontrola integrity

Kontrola integrity dat je jedno ze základních opatření jakékoliv kvalitní správy dat. V oblasti digitální archivace do této oblasti spadají dvě specifické operace: identifikace a validace formátu (21). Identifikace znamená jednoznačné určení formátu, validace stupeň shody (předpokládaného formátu) s jeho oficiální specifikací (8). Identifikace formátu je cílena na jednoznačné určení formátu, tak, aby se odlišil od jiných formátů. Pokud není užit identifikátor PUID, pak je variantou identifikace uvedení několika údajů současně (název, verze formátu). Důvodů potřeby identifikace je několik. Jedním je, aby s daným formátem mohla být jednoznačně propojena rizika a dále též informace o širším prostředí (nástroje potřebné na produkci, konverzi apod.), například odkazem do formátového registru. Dalším důvodem je, aby repozitář měl možnost jednoznačné deklarace požadovaných formátů a jejich kontroly. Identifikace je založena na prostém faktu, že přípona souboru není jedinečný identifikátor a její užití není nijak standardizováno. Identifikace se v praxi děje automatizovaně prostřednictvím speciálních nástrojů. Je třeba poznamenat, že existují různě definované způsoby, resp. úrovně validace, které jsou odvozeny od samotných nástrojů, které validaci provádějí. JHOVE (hlavní nástroj pro validaci v současnosti) rozlišuje již 3 úrovně validace: správná strukturovanost (well-formedness) znamená splnění syntaktických požadavků formátu, validita (validity) splnění sémantických požadavků a konzistentnost (consistency) soulad s externími požadavky (viz profil pro JP2). Zatímco JHOVE je určen k validaci více formátů, novější nástroj jpylyzer je určen výhradě pro validaci formátu JP2 (a také definuje pouze jednu úroveň validace: validní / nevalidní). Výstupy identifikace i validace by měly v optimálním případě být zapsány do metadat, a to včetně informací o užitém validátoru a jeho verzi. Důvodem je existence různých nástrojů a možnost dodatečné (budoucí) kontroly, resp. úvahy o důvěryhodnosti těchto nástrojů, a z toho plynoucí možnost později rozhodnout, že tyto operace bude repozitáře v budoucnosti opakovat. Identifikaci, resp. validaci mohou vykonávat vkladatel i repozitář (dvojí kontrola), nebo jen repozitář; záleží na dané dohodě o dodávání dat. Repozitář by měl optimálně provádět vždy určitou úroveň validace a explicitně uvádět, jaké validace provádí, resp. neprovádí.

K identifikaci je v NDK užíván zmíněný nástroj DROID. Plánuje se podrobnější testování novějších nástrojů pro identifikaci využívajících též registr PRONOM (FIDO

a Siegfried). Validace datového formátu JP2 se v současnosti při příjmu do LTP úložiště provádí pouze na vybraném vzorku. Je využit nástroj jpylyzer, validace konzistentnosti se provádí na základě srovnání výstupů jpylyzeru s předepsanými profily pro JP2. V případě nesrovnalostí se zapojí JHOVE jako druhý kontrolní validátor. V softwarové komponentě systému LTP úložiště jsou nastaveny povolené datové formáty podle identifikátoru PUID. Pokud PUID neodpovídá povoleným formátům, systém jej nepřijme. Tento postup se spoléhá na vkladatele a jeho závazek dodávat metadata v předepsané podobě a korektně vyplněná. Výše uvedený postup není optimální. Jednak nejsou prováděny kompletní validace všech příchozích dat, jednak princip důvěry ve vkladatele je sice legitimní, ale validace předepsaných formátů samotným repozitářem je nepochybně jistější přístup. Z těchto důvodů nyní v LTP úložišti probíhají práce na vytvoření komplexního validátoru. Tento validátor bude provádět kontrolu, zda obrazové soubory a metadata odpovídají předepsaným profilům. V případě metadat půjde nejen o kontrolu struktury, ale též vybraných obsahových položek. U obrazových dat bude kontrolována validita a konzistentnost a jejich zobrazitelnost v aplikaci. Validátor bude využívat několika nástrojů. Nástroje JHOVE a jpylyzer budou užity pro kontrolu validity, jpylyzer též pro kontrolu konzistentnosti. Kontrola zobrazitelnosti je funkce, která jde nad rámec výše uvedených formátových nástrojů. V minulosti se stávalo, že ačkoliv byly některé obrazové soubory validní (podle jpylyzeru i JHOVE), nebylo možné je zobrazit v našem prezentačním systému, což lze považovat za výrazné riziko. Zobrazitelnost obrazu by podle našich testů měly odhalit nástroje Image Magick a Kakadu (22). Oba nástroje dokázaly určit problémové obrazy. Image Magick identifikoval všechny obrazové soubory, které Kakadu (jako komponenta našeho prezentačního systému) nedokázal zpracovat. Důvodem, proč je nástroj Kakadu také začleněn do validátoru, je zvýšení kontroly a právě jeho užití v naší praxi. Hotový validátor bude zpřístupněn i ostatním knihovnám. Knihovny, které budou mít licenci na Kakadu, budou mít k dispozici verzi obsahující všechny součásti validátoru. Knihovny bez této licence budou mít k dispozici verzi validátor bez nástroje Kakadu.

8 Závěr

Formátová strategie je jednou z nejdůležitějších součástí strategie digitální archivace. S jistým zjednodušením lze říci, že uložením do archivačního formátu (a jeho kontrolou) odpadá repozitáři po následujících několik let velká část problémů – pokud je ovšem takový formát pro danou oblast vůbec k dispozici. V případě rastrových dat tomu tak naštěstí je – existuje několik vhodných archivačních formátů. Každý formát,

i archivační, ovšem představuje určité riziko, byť sebemenší. Z tohoto důvodu je nezbytné pravidelné sledování vývoje technologií. Formát JP2, v současnosti jediný archivační formát pro rastrová data přijímaný do LTP, se od dob jeho prvního užití v projektu NDK celosvětově ještě více rozšířil. Dokonce i americké paměťové instituce (které byly dlouho k tomuto formátu skeptické a za de facto standard pro archivační formát platil nekomprimovaný TIFF) začaly v posledních letech ve větší míře zvažovat jeho užití. Obecná strategie přejímat mezinárodní postupy a doporučení je základem naší praxe. Je klíčové využívat globální metadatové standardy a také široce užívané nástroje a služby pro práci s formáty. Standardizace a široká užívanost jsou zárukou interoperability a také vyšší pravděpodobnosti dalšího vývoje a odhalování chyb a jejich oprav (princip „konzistentně špatně“). Nedílnou součástí naší praxe je testování nástrojů a vlastní vývoj. Současné postupy LTP úložiště vyžadují řadu zlepšení, k nimž z plánovaných akcí nejvíce přispěje nový komplexní validátor, který mj. zajistí plně automatizovanou, celoplošnou (na všech datech) a kompletní validaci (tj. včetně konzistentnosti) obrazových dat, která nám dosud chyběla a která je spolu s výběrem vhodného archivačního formátu klíčovým prvkem digitální archivace v úvodní fázi řízení životního cyklu dokumentů, ve které se nachází většina paměťových institucí na světě.

9 Citované zdroje

1. 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.
2. ČSN ISO 16363. *Systémy pro přenos dat a informací z kosmického prostoru – Audit a certifikace důvěryhodných digitálních* úložišť. Praha: Úřad pro technickou normalizaci, metrologii a státní zkušebnictví, 2014. 53 s.
3. Cubr, Ladislav, et al. Srovnání vybraných národních identifikačních systémů užívajících identifikátory URN:NBN. *ProInflow* [online]. 2016, Vol. 8 (No. 1) [cit. 2016-10-14]. ISSN 1804–2406. Dostupný z WWW: <http://www.phil.muni.cz/journals/index.php/proinflow/article/view/1220>
4. Kejser, Ulla Bøgvad; Nielsen, Anders; Thirifays, Alex. Cost Aspects of Ingest and Normalization. In: *IPRES 2011 – 8th International Conference on Preservation of Digital Objects* [online]. Singapur : National Library Board Singapore, 2011, s. 107-115 [cit. 2016-10-14]. ISBN 978-981-07-0441-4. Dostupný z WWW: <https://phaidra.univie.ac.at/detail_object/o:294293>.

5. *Technical Guidelines for Digital Cultural Content Creation Programmes: Version 2.0* [online]. MINERVA eC Project, 2008 [cit. 2016-10-14]. Dostupný z WWW: <http://www.minervaeurope.org/publications/MINERVA%20TG%202.0.pdf>.

6. FADGI Still Image Working Group. *Raster Still Images for Digitization: A Comparison of File Formats : Part 2. Detailed Matrix (multi-page)* [online]. Federal Agencies Digitization Guidelines Initiative, 2014 [cit. 2016-10-14]. Dostupný z WWW: <http://www.digitizationguidelines.gov/guidelines/FADGI_RasterFormatCompare_p2_20140417.pdf>

7. *Library of Congress Recommended Formats Statement 2016-2017* [online]. Washington: Library of Congress, 2017. [cit. 2016-10-14]. Dostupný z WWW: <http://www.loc.gov/preservation/resources/rfs/RFS%202016-2017.pdf>.

8. Abrams, Stephen L.; Seaman, David. Towards a global digital format registry. In *World Library and Information Congress : 69th IFLA General Conference and Council, 1-9 August 2003, Berlin* [online]. Berlin : IFLA, 2003 [cit. 2016-10-14]. 10 s. Dostupný z WWW: <http://archive.ifla.org/IV/ifla69/papers/128e-Abrams_Seaman.pdf>.

9. Brown, Adrian. *The PRONOM PUID Scheme : A scheme of persistent unique identifiers for representation information* [online]. The National Archives, 2006. [cit. 2016-10-14]. 9 s. Dostupný z WWW: <http://www.nationalarchives.gov.uk/aboutapps/pronom/pdf/pronom_unique_identifier_scheme.pdf>.

10. Vychodil, Bedřich. JPEG2000 – Aneb nemyslete si, že vás mine! *Knihovna : knihovnická revue*. 2010, roč. 21, č. 2, s. 53-68.

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

12. Buonora, Paolo; Liberati, Franco. A format for digital preservation of images : a study on JPEG 2000 file robustness. *D-Lib magazine* [online]. July/Aug 2008, vol. 14, no. 7/8 [cit. 2016-10-14]. Dostupný z WWW: <http://www.dlib.org/dlib/july08/buonora/07buonora.html>.

13. Buckley, Robert; Sam, Roger. *JPEG 2000 Profile for the National Digital Newspaper Program* [online]. Washington : Library of Congress, 2012 [cit. 2016-10-14]. 24 s. Dostupný z WWW: <http://www.loc.gov/ndnp/guidelines/docs/NDNP_JP2HistNewsProfile.pdf>

14. Buckley, Robert. *JPEG 2000 as a Preservation and Access Format for the Wellcome Trust Digital Library* [online]. London : KDCS Digital Consultancy, 2009 [cit. 2016-10-14]. 17 s. Dostupný z WWW: <http://wellcomelibrary.org/assets/wtx056572.pdf>.

15. Palmer, William; May, Peter; Cliff, Peter. An Analysis of Contemporary JPEG2000 Codecs for Image Format Migration. In *Proceedings of the 10th International Con-*

ference on Preservation of Digital Objects [online]. Lisbon : Biblioteca Nacional de Portugal, 2013, s. 197-202 [cit. 2016-10-14]. Dostupný z WWW: <http://purl.pt/24107/1/iPres2013_PDF/iPres2013-Proceedings.pdf>.

16. Svoboda, Tomáš, et al. *Elektronické publikace v Národní knihovně ČR* [online]. 1. vydání. Praha: Národní knihovna České republiky, 2015 [cit. 2016-10-14]. ISBN 978-80-7050-654-7. Dostupný z WWW: <https://drive.google.com/file/d/0B46gpfbHV70tR28wSzYyUmxqOHc/view>

17. Ostráková, Natalie. *Analýza archivačních formátů pro zvukové dokumenty užívaných v zahraničních institucích*. Praha : Národní knihovna ČR, 2016. Interní dokument.

18. Sanz, Pascal. Development of electronic periodicals at the Bibliothèque nationale de France : digitisation of French daily newspapers from Mid 19th Century to 1944. In *World Library and Information Congress : 71th IFLA General Conference and Council : "Libraries – A voyage of discovery", August 14th – 18th 2005, Oslo, Norway* [online]. 2005 [cit. 2016-10-14]. Dostupný z WWW: <http://archive.ifla.org/IV/ifla71/papers/141e_trans-Sanz.pdf>.

19. *JPEG 2000 profiles – examples from a range of institutions (footnotes on reverse)* [online]. Digital Preservation Coalition [cit. 2016-10-14]. Dostupný z WWW: <http://www.dpconline.org/component/docman/doc_download/529-jp-2knov2010parametercomparisonchart>.

20. ZENG, Marcia Lei; 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.

21. *PREMIS data dictionary for preservation metadata version 3.0* [online]. Washington (DC, USA) : Library of Congress, 2012. viii, 273 s. [cit. 2016-10-14]. Dostupný z WWW: <http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf>.

22. Ostráková, Natalie. *Test validace a zobrazitelnosti formátu JPEG 2000*. Praha : Národní knihovna ČR, 2016. Interní dokument.

23. Center for Research Libraries; OCLC. *Trustworthy Repositories Audit & Certification : Criteria and Checklist* [online]. Version 1.0. February 2007. Chicago : Center for Research Libraries, 2007 [cit. 2016-10-14]. Dostupný z WWW: <http://www.crl.edu/PDF/trac.pdf>.

File formats for audiovisual preservation: How to choose?

Peter Bubestinger, Osterreichische Mediathek

Abstract:
Every memory institution dealing with audiovisual material must face these questions sooner or later:

- Which file format shall we store our collection in?
- Which video container?
- Which video codec?
- Lossy or lossless encoding?
- Which format can still be read in the future?
- Which file formats are already part of our digital archive?
- Should we replace them, or can we keep them?

No matter whom you ask, you will get different answers. The answers might be correct, but they they might not be the right solution for your use-cases.
There is no evergreen format. There is no one-size-fits-all solution.
This talk addresses what to look out for when choosing a file format: Now and in the future.

1. Introduction

Which properties should an archivist look out for when choosing a format for long-term preservation of audiovisual materials?

When asking other professionals in that domain, they often refer to a format which is promoted and supported as industry standard. Or one is simply choosing to use what others are using for the job.

Today we are living in a time where technology is becoming more and more complex, more intransparent. Especially digital video is a challenge for preservation, since it combines all things to be considered for preserving images, audio and metadata. Additionally, audiovisual material requires vast amounts of storage capacity and data bandwidth. Non-trivial, even with today's existing hardware.

1.1 Different institutions, different use-cases

In difference to digital audio formats, there is currently no "one size fits all" format to be used for digital video. Before we continue, it is important that the reader is aware that there are different use-cases and needs of professional institutions dealing with audiovisual material.

The main ones are:

- production
- broadcast
- preservation

Production and broadcast institutions have their own collections, but their priority are things like retrieval, archive-to-air times and editing. Preservation of their material might of course be in their interest, but it is not their main objective. This should be kept in mind, as it means that solutions designed for production or broadcasting, might not necessarily be the best choice for preservation.

In this paper, we will focus on the use-cases and needs of professional preservation of audiovisual material. Mainly video, but the basic principles can be applied to film and possibly other media types, too.

1.2 Video on tape: An endangered species

In comparison to many other formats that we archivists try to preserve, video is a different challenge on its own:

- Very short market relevance time, compared to other media, such as audio or film for example.
- Therefore:
  - Replayers discontinued
  - Lack of spare parts and service manuals

    - Even for mainstream formats (e.g. VHS)
- Video professionals mainly in broadcast
- Film is very different than video
- Electric tape layout complex to understand
- and many more...

This mostly applies to analogue video, but also to digital video formats stored on tape, such as DigiBeta or DV for example.

1.3 Digital video primer

Digital video always consists of at least 3 different entities. Let's call this the "digital video trinity":

- Video codec
- Audio codec
- Container

For proper preservation purposes, it is also important to consider a fourth entity:

- Metadata

Metadata is very important for proper long-term preservation. Even if it the metadata is just text, it makes a big difference for accessing and using that data, in which way it is digitally stored. An important decision for storing metadata is whether it is to be stored directly inside a media file (=embedded), or in separate files next to the media files (=sidecar).

It may seem more convenient and "cleaner" to embed the metadata, but there are pros and cons of both methods.

Embedding the metadata into the mediafile adds additional complexity, and limits the number of applications that will be able to properly read or write certain metadata, which also increases dependence on certain implementations of hardware/software – as well as the codecs/containers eventually supporting storing the metadata properly.

Sidecar files often allow easier, technology-neutral access of metadata, but may seem more inconvenient if an application does not automatically support reading them. See the chapter "Try before you buy" for more information and practical tests.

This article will not go into details about pros/cons of metadata formats, because that would justify another paper on its own.

The video container format is often what users see as the file ending of a video file. Popular examples for video container formats are “AVI”, “MOV”, “MKV”, “MXF” or nowadays more and more “MP4”.

Unfortunately, it has become very common to use the name of the container only as answer to the question “which format?”. This leads to a lot of uncertainty, confusion and misunderstandings when speaking about video formats, because it’s actually necessary to always consider all three components when speaking about “a video format”: Videocodec, audiocodec and container.

For example the most popular and commonly agreed upon format for archiving audio is often called “WAV” – or “Wave File”. Yet, “WAV” is merely the container – and the actual audio codec is “Linear PCM uncompressed”. The audio codec inside a WAV could also be something different. Even “MP3” as codec in WAV is legal, according to the container specification.

Using uncompressed PCM as audio codec is perfectly possible and well supported in almost all digital video containers. Therefore, this article will only focus on the remaining 2 entities which are most unclear – and also have the most impact for preservation:

- Video codec
- Container

2. Defining technology-neutral properties for long-term preservation

Technology will always change.
This has a direct effect on availability, support and sustainability of tools to handle collection’s contents. At the end of the day, one must make a choice for technology being available to them today – but one day this technology will be outdated or unavailable. Therefore, it is a good idea to define a list of property-requirements that a format for storing ones collection must have. Independent of which technology provides it.

This makes it possible to apply the same set of requirements whenever a new technology, a new format must be chosen to migrate to. Since there is no evergreen format, migration to a yet-unknown format in the future will be necessary one day.

By defining a proper set of preservation format requirements, obstacles for migrating to future formats can greatly be reduced.

Take written text for example: The "technology" being used over ages changed dramatically.

- Carving in bones, stone, wood, etc.
- Drawing on stone, papyrus, wood, paper, leather, etc.
- And many more...

Yet, the required properties for sustaining the content stayed the same:

- Visually perceivable contrast to store and retrieve the information
- Means of understanding the visual content: symbols / language

These requirements could then also be applied to digital technolog: From basic text files to more complex digital document formats.

This approach of defining the needs for long-term preservation in a technology-neutral way, allows to use this set of properties as demands that shall be met when choosing a format for archiving ones content. The example of written text was chosen to illustrate that this principle can also be applied to other media as well. Not only formats for audiovisual media.

3. Property list

The properties listed in this chapter are based on a list of desired format and technology properties for digitizing and preserving video material. This list was put together by technicians at the "Österreichische Mediathek" [1] , Austria's national audio/video archive.

Their motivation and necessity for writing such a list came up in 2009 as a result of the evaluation phase when considering buying a system for bulk-digitization of their video collection [2]. After talking to, and questioning many other memory institutions as well as broadcasting archives, it turned out that there was no commonly agreed upon method – or formats – for preserving video in digital form.

In order to make it easier to apply and use this list, it is here split into 2 different categories:

1. **Significant properties**
These are the properties which are considered significant in order to maintain the actual content as accurately as possible in a technical way.

2. **Preservation-improving properties**
These are properties that increase the chances for long-term preservation. Some of these properties are there to avoid unnecessary issues that may cause additional efforts or problems when dealing with a format in the future. Such as dealing with format obsolescence or future mass migration.

3.1 Significant properties

- No digital loss
- Resolution independent
- Aspect ratio preserving
- Colors as native as possible

3.2 Preservation-improving properties

- Handleable data amount
- Non-proprietary
- Hardware independent
- Avoiding unnecessary complexity

4. Properties in detail

4.1 Significant properties

4.1.1 No digital loss

With analogue material, even high-quality copies were a degradation compared to the original. One of the major benefits of digitization is the fact that the content can be

copied infinitely without any loss. In the audio domain for example, it is already common to record and edit recordings in a digitally lossless way. Non-professional, consumer devices use lossy compression for recording in order to save space. Yet, in professional environments it would be considered inacceptable to record or edit audio material using lossy formats, such as MP3 or MP4, for example. Regardless how good it may sound to the listener.

For audio material archiving, it is therefore mentioned as requirement in the technical guidelines of "IASA-TC03" [3, p.8], that only uncompressed data, or lossless compression is acceptable for proper long-term preservation.

For digital video materialthe situation is currently far from perfect:

The ones who started dealing with preservation of digital video material were mostly broadcast and production. The reason for that is, that they simply have way larger budgets than preservation archives. Therefore, the market focused on delivering solutions tailored towards the needs of their main customers.

As stated above, preservation-needs of broadcast/production are not necessarily the same preservation-needs of archives. Since the data size of digital video is still non-trivial to handle – even by current technology – it became common to use lossy compression formats.

To give you a better understanding of the data sizes to consider when dealing with digital video, here are a few examples:

PAL Standard Definition (SD):

- 720 x 576 pixels resolution
- 25 frames per second (fps)
- 8 bits-per-component YUV
- 4:2:2 chroma subsampling (=16 bits per pixel)

= 720 * 576 * 25 * 16 / 8 / 1024 / 1024 = 19,77 MB per second
= 1,186 GB per minute

PAL High Definition (Full HD):

- 1920 x 1080 pixels resolution
- 25 frames per second (fps)

- 8 bits-per-component YUV
- 4:2:2 chroma subsampling (=16 bits per pixel)

= 1920 * 1080 * 25 * 16 / 8 / 1024 / 1024 = 98,88 MB per second
= 5,8 GB per minute

These are only illustrative data rates for the image data only.
Assuming SDI standard for audio as example, with 48 kHz and 24 bits resolution, that is another 8,2 MB per minute to add for each audio channel present.

One must also consider that additional space might be required for audio tracks, higher framerate (e.g. 50fps), less chroma-subsampling (e.g. 4:4:4) and higher bits-per-component (bpc) sample depth. For film at resolutions of 2k at 14bpc and more, you can imagine the impacts not only on storage, but also CPU processing and network performance requirements.

4.1.2 Resolution independent

Although there are are numerous standard resolutions common for certain video formats, it is desirable to have a preservation format that is able to store arbitrary resolutions. This makes it possible to store any material as natively as possible without the need to resample the image data.

It also allows for using the same format for future resolutions not common at the day of choosing the format. In practice there are cases where the video format definition is not limited to certain resolutions, but the actual implementation of a hardware or software might be.

4.1.3 Aspect ratio preserving

Every image has a so called “aspect ratio”. This is the ratio of its width to its height. There are different aspect ratios to consider when dealing with video:

SAR: Storage Aspect Ratio
The aspect ratio of the image actually stored per image.
Example: 720 x 576 pixels for SD PAL.

PAR: Pixel Aspect Ratio
The aspect ratio of the pixels of the display unit.

Example: Computer screens have square pixels, whereas analogue TVs often had rectangular pixels.

DAR: Display Aspect Ratio
This is the most popular – and most important aspect ratio for the viewer. It defines in which aspect ratio the final image is to be displayed.

Common DAR for video are “4:3” or “16:9” for example.

There are technial relations between all 3 of these aspect ratios. The image of audiovisual recordings is not always stored in the same aspect ratio as it is to be displayed in. It is best practice to store the image as natively as it was on the source.

For example, when recording 16:9 on a DigiBeta, it is stored anamorphic.

This means, that the SAR is 720x576 (=5:4 ratio), but the DAR is to be 16:9. Therefore in this example the image captured correctly “as-is”, would be 5:4 aspect ratio – and therefore distorted / squished to the user’s eye.

By the way, also 4:3 material is actually stored in 5:4 SAR. This is less visible distortion, but it originates from the non-square pixels of CRT screens.

This must be considered by an archivist when choosing how to capture and preserve their collection.

A recording system or file format for digital video must therefore be able to capture the original SAR without any interference or “correction”, while being able to resize the material later on to its proper DAR for research and access copies, for example.

4.1.4 Colors as native as possible

When dealing with images – especially color images – there are different things to watch out for in order to avoid silent modification of the actual color information.
For video colors, important factors are:

Colorspace
Analogue CRTs, as well as PC screens and digital TVs display color dots in Red Green Blue: RGB

This color model is the most native for electronic equipment to record and display.

Yet, for legacy reasons related to introducing color later onto the black-and-white TV infrastructure in the past, a more complex color model was introduced:

Most analogue video – and for legacy and compression reasons even digital video – is a color space commonly referred to as "YUV".

Actually, there are slight differences between the analogue version and the digital one (YCbCr), but important to note here is that it is fundamentally different to RGB

Chroma subsampling
Also originating from the early days of color television being backwards compatible with black-and-white, a form of analogue signal compression was used.

This form of compression is called "subsampling" and is based on the fact that our human perception can be tricked by using full resolution for black-and-white, but less resolution for the color information.

It is usually written in the following form:
- 4:4:4
- 4:2:2
- 4:2:0
- etc...

The important reason to mention it here is, that not all audiovisual media, signal chains or file formats, use and support the same subsampling.

For example, analogue video on VHS is "4:2:2", default on DVD or digital television is "4:2:0" – and Digital Video tapes (DV) for example are "4:1:0".

Please refer to the "Chroma Subsampling" article on Wikipedia [4] for additional information on this topic.
Knowing that almost any video is silently and automatically cross-converted between the colorspace and subsampling in which it is stored, it is still important to consider

these properties when choosing a file format to store the material in. As original and as close to the source as possible.

4.2 Preservation-improving properties

4.2.1 Handleable data amount

Given the huge amounts of data size of audiovisual material, even if it is just video and not even film, several entities of an institution's workflow must be considered.
There is no sense in choosing a format that one cannot handle given the current technology- and/or budget-limitations.

The main entities to be considered are:

- storage size
- network speed
- disk speed
- data bus speed (RAM, etc)

When talking about digital video preservation formats, the most prominent factor mentioned are often the storage costs.

Although storage size is one of the most noticeable cost factors, it is also important to consider the impact of these data sizes on other infrastructure requirements within an institution, too.

Even if a regular office-grade PC is able to playback video with Full-HD resolution smoothly in realtime over a limited bandwidth, such as an Internet connection for example, the same hardware might stutter when trying to play the same video in an uncompressed format.

Therefore, one might want to go through the following checklist:

- How much storage space is needed for 1 copy?
- Plus: At least 1 backup?
- How much data throughput (MB/s) is required for smooth real-time playback?
- Which network speed do I need for this (1GB, 10GB, etc)?
- Do I have enough free network bandwidth for storing daily ingest + backups?
- How many concurrent users do I have, accessing the material in-house?

NOTE: One very important aspect though is, that one must not forget that preservation properties and quality should not suffer. Even if this means additional costs. It is also so, that the most expensive factor in digitization are staff costs. Therefore, it is desirable to do the ingest work only once and as good as possible.

A very common misconception is to use high-quality lossy compression, although this is, again referring to TC03 guidelines for audio, not recommended for long-term preservation.

Regardless if it might look "good enough" for now:

Looking at video encodings done in the recent past of only a few years ago on today's equipment, one can immediately see the impact of the choice for smaller size back then. Even though it was the best available technical solution existing back then.

On the other hand, using lossless compression formats the gain in size due to compression might allow smaller storage (or more backups), as well as using existing network infrastructure for the same amount of playback hours and concurrent users.

4.2.2 Open Source / Non proprietary

The common status quo: proprietary
At the time of this writing, the default of implementations available for dealing with audiovisual material are proprietary.

This means that an end-user does not get information about the inner workings of the equipment they use and require for handling their own material. If one is using proprietary technology or formats, there is a great dependence on the good-will and market interests of that manufacturer.

This might often not really be noticeable until one wants a feature which has too little market-relevance, one wants to migrate to another technology – or another product from another vendor. Or simply the original manufacturer goes out of business.

Such a dependency is called "vendor lock-in".

Although vendor lock-in is not a new concept at all, the nature of older technology – even electromechanical – allowed users to use, reverse-engineer and modify such

equipment in ways that did not need the explicit aid or consensus of the original manufacturer.

This is very different with digital technology and modern electronics:

In the past, skilled engineers employed in archives were able to understand and even fix problems, by applying common-sense and their electromechanical understanding of things. Nowadays, try to find out why a digital video works in one application, and fails to render correctly in another. As we can see in practice as well as in other areas of preservation, archivists need to be able to maintain their equipment independent of whether a certain technology or format is still actively available on the market. Just take a look at what is necessary and current practice for keeping old equipment alive and working for reproduction of media such as audio tapes, vinyl records or film – to mention just a few prominent ones.

In the past it was common for professional equipment to come with so called "Service Manuals". These included schematics and often detailed information about the machines themselves, as well as how to modify, adapt and repair them. Taking a look at all fields of long-term preservation, in order to preserve and access old material, archivists are still "hacking" together clever solutions to challenging problems on a daily basis. With great success.

The same requirements and necessities are to be applied to electronic and digital equipment.

For some reason, it has become increasingly common by vendors to reduce and withhold technical information from their customers. Up to a point where they become even unfriendly if one asks them for specification or implementation details necessary to access or migrate collection material. Choice for technology used in archives is getting increasingly linked to the product market lifetime: An average of 3 to 5 years. I hope that the reader will agree that memory institutions define "long-term" as a much more longer period that that.

It is therefore obvious, that this dependency must be kept as little as possible by archives in order to be able to fulfill their task of proper long-term preservation.

A practical solution: Free and Open Source

Imagine one could archive not only existing equipment, such as replayers for example, but also the schematics and building components required to use, study, share and improve their equipment as it fits ones needs.

For software, this is possible already today.

The license definition of "Free and Open Source Software" (FOSS) states that certain conditions must be provided to the end-user at all times [16].

This is defined by the rights to:

- use
- study
- share
- improve

Every computer program is built out of its "source code". This source code is nothing else than just written text, interpreted by a computer. Having a copy of the source code of the tools used to create or open digital formats, makes it possible for developers to adapt it to any future technology. Even if not known yet.

This is the equivalent to what has always been done in the past to make content stored on outdated technologies available until today. Therefore, using FOSS for preservation purposes, counteracts issues such as format obsolescence or vendor dependency.

Not by chance, but by license definition.

It should be noted that a misconception currently common is to mistake "Free and Open Source Software" with "Freeware". Although it is often so, that many Open Source programs are freely (gratis) available, this is not mandatory. Therefore Open Source is not the opposite of "commercial", but merely the opposite of proprietary, closed source.

The quality of any software application (FOSS or proprietary) is independent of its price and license – but in case of Open Source, the end user is in control.

Although archives are most often not equipped with IT development staff, FOSS still enables institutions and individuals to combine their resources and have software tailored to their needs. Without the necessity to reinvent the wheel or start from scratch, because it is very likely that other existing FOSS code can be based on, or simply incorporated in their new solution.

In recent years, special tools for needs of archivists dealing with digital video have been released as Free Software/Open Source.

These were often using existing code from other software projects, or adding new features that had too little market-relevance, but are invaluable for archiving.

Some popular examples are:

- QCTools: Digital quality control (QC) for recorded analogue video tapes [6].
  Originally funded by Bay Area Video Coalition.
- MediaInfo: A tool for displaying of the most relevant technical and tag data for video and audio files [7].
  Funded by several audiovisual institutions and archives – as well as the European Broadcasting Union (EBU).
- FFV1: A lossless video compression codec [8].
- Ffmpeg/Libav/FFmbc: Complete, cross-platform solutions to record, convert and stream audio and video [9,10] – as well as a "version customized for broadcast and professional usage" [11].
- MediaConch: Implementation checker, policy checker, & reporter for Matroska, FFV1 & PCM [12].
  Originally funded by European Union project "PREFORMA" [13].

It is even so, that hardware that can be run with FOSS rather than its proprietary firmware can be used (a) longer than its original intended market-lifetime, and (b) can be adapted to future conditions, or use-cases it was never designed for.

A popular example from a completely different domain would be the wireless router "WRT54GL" from Linksys [5]: It is the only router which is still available on the market for more than 12 years. This is due to the fact that it can be run on a FOSS firmware.

As these real-world examples show, especially memory institutions can profit from Free and Open Source solutions.

So, using FOSS has the following benefits by license-definition:

1. No vendor lock-in.
2. No black-box technology.
3. Ability to use/study/share/improve their tools.
4. Independence of market interests.
5. No format obsolescence.
6. Future proof by archiving the source code.

If an existing FOSS application does not fit their needs, they have the option to hire developers – and even pool ressources with others to save money.

4.2.3 Hardware independent

Basically, the reasons for trying to avoid hardware dependency are similar, but not the same, as to the ones avoiding manufacturer dependency, mentioned above.

It is still common for manufacturers to offer archiving systems that use certain hardware to generate the actual codec bits stored in the video files. This sometimes leads to the case where one requires certain hardware to be able to decode the video files properly. Even if the format of choice is considered to be an official Standard (like ISO, SMPTE, etc), the actual implementation (in this case in hardware) might create files which render differently – or not at all – on other systems.

This is not something evil that vendors do, but everyone makes mistakes – and you would not want to be relying solely on a piece of market-niche hardware which is usually very expensive, and secondly hard to get or impossible to fix a few years after you bought the system.

In order to spare yourself this trouble, it is better to avoid hardware dependency wherever possible.

4.2.4 Avoiding unnecessary complexity

There is a nice definition of a so called “Minimalistic Data Format” [14] for file formats:

- As simple as possible.
- As complicated as necessary.

Applying these criteria to a digital file format greatly reduce the possible problems. It is actually quite obvious, and analoguous to experiences seen with older media.

Take an old telephone aparatus for example, and compare it to a modern smartphone: Which one lasts longer – and is easier to understand, use or repair?

Of course the old telephone cannot perform a fraction of the functionalities a smartphone does, but it does one thing – and it does it well.

Applying this principle to digital file formats (of any kind) increases your chances of preserving the actual contents.

Trying to create a “Jack of all trades” might in the end lead to a “Master of none”.

And with digital formats, this is the number one cause for interoperability issues, because everyone just implements a subset of the whole set of possibilities a format can offer – and therefore creating incompatible files.

For long-term preservation, the principle of “As simple as possible – and as complicated as necessary” has proven to be even more important in the ever increasing complexity and feature-war of file formats.

5. Try before you buy

5.1 Concept

Regardless which software or hardware solution one is choosing, it is good practice to “try before you buy”. Even if a file format is an official standard (e.g. defined by DIN, ISO, SMPTE, etc) – there are still variations in the actual implementations. These variations differ from implementation to implementation and often cause interoperability issues.

As mentioned above, the more complex a format specification is, the more “interpretation variations” may exist.

A rule of thumb is:

1. The smaller the userbase, the smaller the market.
2. The smaller the market, the smaller the number of different manufacturers providing a certain technology.
3. Only the popular use-cases are stable, due to less interest in testing/maintaining the other cases.

Professional video and film is a very special domain, often associated with high costs, due to it being a niche market. In order to avoid interoperability issues, or migration problems later on, it can save you a lot of time, nerves – and money – to perform the following tests before you agree to any file format technology. Yet, FOSS allows to go beyond black-box testing, since detected issues can often be fixed, rather than being forced to use workarounds.

The following prerequisites apply to all tests mentioned here:

- Only use hardware/applications not originating from the same manufacturer.
- Due to above explained reasons, use FOSS and Open Hardware for testing wherever possible.

5.2 Testset

Before you can begin testing, ask to get representative example copies in the format to be tested.

These files must have been created using the tools you are planning to use to create your preservation archive master files.

5.2.1 Playback on other equipment/appliction:

Can the video be played back properly on equipment or software applications from other manufacturers?

5.2.2 Hardware dependency:

Is proprietary hardware required to properly play it back (e.g. decoder card)?

The general aspects of hardware dependency were already addressed in the previous chapter. For encoding/decoding of digital video it is sometimes necessary to make use of specialized chip-implementations to perform the actions in real-time. Yet, one can simply try to convert the files to another format which can be played back.

If dealing with a lossless format, one can use the method of framemd5 check [15] to verify that the audio/video data was converted bit-proof from format A to format B.

If performance does not allow playback of the archive master format, it is also possible to use another, less ressource demanding codec as target format for the test.

5.2.3 Access metadata:

Which metadata can you access/extract using applications from other manufacturers? Try to read and interpret the metadata provided in the given format. It does not really matter whether the metadata is embedded or sidecar, yet the challenges and troubles to expect with embedded metadata are far greater. Please keep in mind the machine-readability of the metadata. You might be misled by just looking at the metadata, because machines handle it differently than humans. If possible, try to access and match the metadata from the sample source to another target format and see if it works.

5.2.4 Transcode to lossless/uncompressed:

Can you transcode the file to lossless/uncompressed encoding using Free Software / Open Source tools?

If this works flawlessly, you can archive the FOSS tools (and its source code) you've used in your tests. As explained above, this greatly improves the chances of being able to get out of this format in the future. It also increases your chances of being able to perform a fully-automated mass-transcoding for other use cases, such as access copies on the web, for example.

5.2.5 Audio/video synchronicity:

In all your tests, observe if the audio still matches up with the video – throughout the duration of the sample.

Therefore it is good to have samples of a longer duration, since some A/V sync issues only become apparent after a while, due to sync-errors adding up.

If audio/video stay perfectly in sync for the whole duration, perform the same test on a transcoded copy of your sample.

5.2.6 Format specification and details:

Try to get the format specification and/or source code of the implementation.

Ask the manufacturer for a copy of the format specification that they used. Even if the format is an open standard, there might be multiple revisions – and not all specifications are freely available. These specification papers (e.g. ISO) often cost a few hundred Euros per paper.

Additionally, ask them which parts of the standard they have implemented. It is often so, that only a subset of features are relevant for a certain product.

You will in the reaction of the manufacturer if they are willing to cooperate with you in your interest of preservation. If they are reluctant on handing you information that is relevant for you to do your job, your chances are that it will not be better once you are more depending on their implementation.

If possible, try to get a copy of the source code used for the actual implementation you are going to use.

Although FOSS should be preferred (due to reasons mentioned in great detail in the above chapter), it is still possible to get source code from proprietary vendors under non-disclosure agreements.

Having the source code greatly improves your chances to be able to open the format in the future.

5.2.7 Lossless editing:

Can you edit and export the file in applications from other manufacturers – without any digital loss?

Many applications used even in professional studios are performing silent conversions to the audiovisual data in the background.

Try if you are able to load the sample file in a video editing suite, and export an excerpt without any modification to the actual audiovisual data.

If it is not possible to verify losslessness of this process, you cannot be sure what was changed – and if there are generation loss issues to be expected.

6. Conclusion

For additional information, one might also consider reading the following two articles which were written based on practical experiences with digitizing, preserving and handling digital video at the Österreichische Mediathek:

- The archivist's video codec and container FAQ [17]
- Comparing video codecs and containers for archives [18]

There is no silver bullet for choosing the right file formats for preserving audiovisual material in digital form.

Yet, I hope that this paper provides you with information to understand a bit more what to look out for – and hopefully be able to ask the right questions to your suppliers.

7. References

[1]
Website: "Österreichische Mediathek"
http://www.mediathek.at/

[2]
DVA-Profession Documentation: "Project Motivation"
Hermann Lewetz
August 2011
http://dva-profession.mediathek.at/fileadmin/MEDIASERVER/dva-profession-html/documentation/projekt-motivation/index.html

[3]
IASA-TC 03: "The Safeguarding of the Audio Heritage: Ethics, Principles and Preservation Strategy"
International Association of Sound and Audiovisual Archives, Technical Committee
December 2005
http://www.iasa-web.org/sites/default/files/downloads/publications/TC03_English.pdf

[4]
Wikipedia: "Chroma subsampling"
https://en.wikipedia.org/wiki/Chroma_subsampling

[5]
“Linksys WRT54GL Wireless-G Wireless Router”
Linksys
http://www.linksys.com/us/p/P-WRT54GL/

[6]
Website: QCTools
Bay Area Video Coalition
https://www.bavc.org/preserve-media/preservation-tools/qctools

[7]
Website: MediaInfo
Jérôme Martinez, MediaArea
https://mediaarea.net/en/MediaInfo

[8]
Wikipedia: “FFV1”
https://en.wikipedia.org/wiki/FFV1

[9]
Website: “FFmpeg”
http://ffmpeg.org/

[10]
Website: “Libav”
https://libav.org/

[11]
Website: “FFmbc: FFmpeg customized for broadcast and professional usage”
https://github.com/bcoudurier/FFmbc

[12]
Website: “MediaConch”
https://mediaarea.net/MediaConch/

[13]
Website “PREFORMA, PREservation FORMATs for culture information/e-archives”
http://preforma-project.eu/

[14]
“Minimalistic Data Format”
Free Software Foundation Europe
https://fsfe.org/activities/os/minimalisticstandards.en.html

[15]
“Video per-frame integrity check – verifying losslessness”
Peter Bubestinger
January 6th, 2011
http://www.das-werkstatt.com/forum/werkstatt/viewtopic.php?f=7&t=1836

[16]
“What is Free Software?”
Free Software Foundation Europe
https://fsfe.org/about/basics/freesoftware.en.html

[17]
“The archivist’s video codec and container FAQ”
Peter Bubestinger, Hermann Lewetz, Marion Jaks
September 23rd, 2016
http://download.das-werkstatt.com/pb/mthk/info/video/FAQ-digital_video_archiving.html

[18]
“Comparing video codecs and containers for archives”
Peter Bubestinger, Hermann Lewetz, Marion Jaks
August 13th, 2015
http://download.das-werkstatt.com/pb/mthk/info/video/comparison_video_codecs_containers.html

Webový archívny formát WARC

Andrej Bizík, Univerzitná knižnica v Bratislave

Abstrakt
Za posledné roky sa pamäťové organizácie snažia nájsť najvhodnejší spôsob, ako sledovať zmeny a zhromažďovať webové stránky a webový obsah, ktorý sa mení každý deň. Je dôležité, aby formáty umožňovali v jednom súbore jednoducho a bezpečne uschovať veľmi veľký počet dátových objektov rôznych formátov a typov pre skladovanie, riadenie a výmenu dát. Preto bol vytvorený medzinárodný štandard ISO, opisujúci webový archívny formát WARC. Článok obsahuje krátku históriu, základné údaje a stručný popis súboru WARC. Zároveň sa zaoberá popisom fungovania webového archívu slovacikálneho obsahu, ktorý začala realizovať Univerzitná knižnica v Bratislave v roku 2015 ako národný projekt „Digitálne pramene – webharvesting a archivácia e-Born obsahu“. Pilotná prevádzka projektu skončila k 31.12.2015, od januára 2016 na ňu nadväzuje prvý rok udržateľnosti. Pre potreby realizácie projektu vzniklo v rámci Národnej agentúry ISSN oddelenie Depozit Digitálne Pramene (DDP), ktoré sa stalo súčasťou rutinnej prevádzky. Za toto obdobie sa uskutočnilo niekoľko archívnych zberov, uložených vo formáte WARC, ktorých počet a veľkosť sú vyhodnotené v závere.

Kľúčové slová: WARC formát, archivácia webu, webový archív

História

Súbor ARC (Arc File Format) začala interne používať organizácia Internet Archive na záznam postupností obsahu, so stručným popisom zozbieraných súborov. Archívny súbor ARC zhromažďuje údaje vo veľkých súhrnných súboroch pre jednoduché uchovanie v konvenčnom systéme súborov. Formát ARC riadi veľké množstvo malých súborov vo veľkých systémových súboroch. Bol navrhnutý tak, aby súhrnné objekty boli identifikované bez použitia súboru. Zároveň ukladá súbory načítané pomocou rôznych sieťových protokolov a po prvom zápise je integrita súboru nezávislá na následnom

obsahu. Pre načítanie objektov z konkrétneho archívneho súboru je dôležité udržať informácie o názvoch súborov, o ich veľkosti a o tom, ako sú navzájom previazané. Indexovanie zaznamenáva spracovávanie súborov a informácie o nich ukladá do databázy pre jednoduché vyhľadávanie, no nesnaží sa štandardizovať formát súborov.

Motivácia pre rozšírenie formátu ARC vzišla z diskusie a skúseností medzinárodného konzorcia Internet Preservation Consortium (IIPC), ktorého členmi sú Internet Archive, národné knižnice a od roku 2015 aj Univerzitná knižnica v Bratislave. Formát WARC je rozšírením formátu ARC, ktorý používa Internet Archive od roku 1996. Formát WARC sa líši od ARC tým, že ponúka nové možnosti, najmä pokiaľ ide o zaznamenávanie hlavičiek HTTP a metadát, pridelenie identifikátora pre každý obsiahnutý súbor, deduplikáciu obsahu a podobne. Formát WARC spravuje štruktúru a ukladá miliardy zdrojov zhromaždených z webu. Formáty WARC a ARC sú dostatočne odlišné, aby ich softvér jednoznačne rozpoznal a správne spracoval oba typy záznamov. Vzhľadom k veľkému množstvu už existujúcich archívnych dát vo formáte ARC je dôležité, aby sa pri prechode do formátu WARC neporušili záznamy.

Medzinárodný štandard ISO 28500: 2009 [1] zdokumentoval súbor vo formáte *WARC*. Norma popisuje formát súboru webového archívu WARC, ktorý ponúka konvencie pre zreťazenie viacerých dátových objektov do jedného dlhého súboru. Formát možno použiť na vytváranie aplikácií pre zber, správu, prístup a archiváciu obsahov stránok [2].

Ide o dostatočne flexibilný formát, ale ani ten nie je dokonalý. Preto sa na workshope konferencie IIPC GA v Stanforde v roku 2015 o nedostatkoch normy prijal návrh WARC Format 1.1 a vznikol projekt o špecifikovaní WARC 1.1 [3]. Projekt eviduje všetky požiadavky workshopu, ako aj navrhované zmeny z rôznych zdrojov. Vzniknutá predloha bude nakoniec odovzdaná medzinárodnej organizácii ISO na kontrolu a schválenie [4].

Typy WARC súborov

Záhlavie súboru WARC, ako aj každý jeho typ, obsahuje záznam o formáte a číslo verzie. V každom zázname sú pomenované polia. Každé uvedené pole sa skladá z názvu, po ktorom nasleduje dvojbodka („:“) a hodnota poľa. Názvy polí sa skladajú z veľkých a malých písmen. WARC súbor obsahuje za názvom poľa záznamové parametre, obsahujúce dôležité informácie (napr. identifikátor záznamu, čas vytvorenia,

dĺžka obsahu, typ obsahu). Každý záznam má uvedený typ v poli WARC-Type. Existuje osem typov záznamov, ktoré rozširujú formát WARC („warcinfo“, „response“, „resource“, „request“, „metadata“, „revisit“, „conversion“, a „continuation“). Polia sú v rôznom poradí hodnôt v kódovaní UTF-8 (8-bitový Unicode Transformation Format) [5]. Tieto typy sú dôležité hlavne pre softvér na čítanie súborov WARC.

„warcinfo“
Typ opisuje informácie o celom súbore, ako dátum vytvorenia, názov a podobne.

V prípade webového archívu obsahuje informácie o iniciátorovi zberu webu, ktorý vytvoril nasledujúce záznamy. Všetky tieto informácie sú voliteľné a pri každom súbore sa môžu líšiť. Základné kontaktné informácie obsahujú meno, názov organizácie, kontaktnú adresu autora, názov a verziu softvéru použitú na vytvorenie súboru, politiku pre rešpektovanie súboru robots.txt na webových stránkach, pracovný názov a IP adresu stroja, ktorý tento WARC vytvoril. Tieto informácie sa nachádzajú zvyčajne raz na začiatku súboru.

```
WARC/1.0
WARC-Type: warcinfo
WARC-Date: 2016-08-30T12:26:14Z
WARC-Filename:
WEB-20160830122614895-00000-12939~worker104.webdepozit.sk~20138.warc.gz
WARC-Record-ID: <urn:uuid:56ff4821-7a40-4757-b585-69d4ac388bf5>
Content-Type: application/warc-fields
Content-Length: 546

software: Heritrix/3.2.0 http://crawler.archive.org
ip: 10.109.33.230
hostname: worker104.webdepozit.sk
format: WARC File Format 1.0
operator: Andrej Bizik, Peter Hausleitner
publisher: Univerzitna kniznica v Bratislave
audience: webdepozit.sk users
isPartOf: basic
description: Archivacia stranok pre dalsie generacie
robots: obey //rešpektovať
http-header-user-agent: Mozilla/5.0 (compatible; heritrix/3.2.0 +
http://www.webdepozit.sk)
http-header-from: admin@webdepozit.sk
```

Obrázok 1: Ukážka typu záznamu warcinfo

„response“
Obsahuje prijatú odpoveď HTTP zo stránky, v prípade dostupnosti aj informáciu o sieťovom protokole. Medzi základné údaje patrí doménové meno, dátum, identifikátor

a kompletná odpoveď. Presný obsah je určený podľa typu záznamu a tiež podľa schémy URI (jednotného identifikátora zdroja) záznamu.

```
WARC/1.0
WARC-Type: response
WARC-Target-URI: dns:consilium.europa.eu
WARC-Date: 2016-08-30T12:26:14Z
WARC-IP-Address: 10.109.22.31
WARC-Record-ID: <urn:uuid:f93ebb05-9c2c-4d47-8785-8071997c12f8>
Content-Type: text/dns
Content-Length: 61

20160830122614
consilium.europa.eu.    86352   IN  A   91.194.202.11
```

Obrázok 2: Záznam typu response

„resource“
Obsahuje zdroj súboru. Zdroj možno špecifikovať podľa viacerých schém, môže odkazovať na úložisko archívu alebo internetu, bez informácií o protokole. Napríklad schéma „http“ alebo „https“ obsahuje záznam o cieľovej adrese URI a schéma „dns“ obsahuje typ obsahu (Content-Type), ktorý obsahuje cieľovú adresu URI (WARC-Target-URI).

```
WARC/1.0
WARC-Type: resource
WARC-Target-URI: file://var/www/htdoc/images/logoc.jpg
WARC-Date: 2006-09-30T16:40:32Z
WARC-Record-ID: <urn:uuid:23200706-de3e-3c61-a131-g65d7fd80cc1>
Content-Type: image/jpeg
WARC-Payload-Digest: sha1:DBXHDRBXLF4OMUZ5DN4JJ2KFUAOB6VK8
WARC-Block-Digest: sha1:DBXHDRBXLF4OMUZ5DN4JJ2KFUAOB6VK8
Content-Length: 1662
```

Obrázok 3: Záznam typu resource

„request“
Záznam obsahuje podrobnosti o úplnej žiadosti v zmysle § 5 HTTP / 1.1 (RFC2616) [6], vrátane informácií o sieťovom protokole. Blok by mal obsahovať správu o žiadosti, teda požiadavku HTTP odoslanú cez sieť, vrátane hlavičky. Pole WARC-IP-adress býva použité na zaznamenanie sieťovej IP adresy. Záznam nešpecifikuje údaje o „https“, ako sú napríklad certifikáty.

```
WARC/1.0
WARC-Type: request
WARC-Target-URI: http://consilium.europa.eu/robots.txt
WARC-Date: 2016-08-30T12:26:15Z
WARC-Concurrent-To: <urn:uuid:cc6acbff-7418-48bb-b46a-63c9dbc81383>
WARC-Record-ID: <urn:uuid:580acff0-683c-4c0a-b3e7-d8243459bb64>
Content-Type: application/http; msgtype=request
Content-Length: 246

GET /robots.txt HTTP/1.0
User-Agent: Mozilla/5.0 (compatible; heritrix/3.2.0 +http://www.webdepozit.sk)
From: admin@webdepozit.sk
Connection: close
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8
Host: consilium.europa.eu
```

Obrázok 4: Záznam typu request

„metadata“

Polia „metadata“ vytvárajú obsah, s cieľom ďalej popísať zozbierané zdroje. Záznam môže odkazovať na iný záznam s informáciou pôvodného alebo transformovaného obsahu. Akýkoľvek počet metadát môže odkazovať na jeden konkrétny záznam. Formáty záznamov sa môžu líšiť a všetky polia sú voliteľné. Pole „via“ konkretizuje stránku, z ktorej bol obsah archivovaný. Čas zberu stránky v milisekundách obsahuje pole „fetchTimeMs“.

```
WARC/1.0
WARC-Type: metadata
WARC-Target-URI: http://consilium.europa.eu/robots.txt
WARC-Date: 2016-08-30T12:26:15Z
WARC-Concurrent-To: <urn:uuid:cc6acbff-7418-48bb-b46a-63c9dbc81383>
WARC-Record-ID: <urn:uuid:f00edaee-c6de-48df-a437-c2bc3552a63b>
Content-Type: application/warc-fields
Content-Length: 310

force-fetch:
via: http://consilium.europa.eu/
hopsFromSeed: P
fetchTimeMs: 186
charsetForLinkExtraction: UTF-8
outlink: http://www.consilium.europa.eu/robots.txt R Location:
outlink: http://consilium.europa.eu/favicon.ico I =INFERRED_MISC
outlink: http://www.consilium.europa.eu/robots.txt L a/@href
```

Obrázok 5: Záznam typu metadata

„revisit“

Nepovinný záznam, ktorý porovnáva zbieraný obsah s už archivovaným obsahom s cieľom nájsť rovnaký obsah, hlavne z dôvodu šetrenia pamäti. Účelom tohto typu záznamu je znížiť ukladanie duplicitného obsahu pri opakovanom načítaní rovnakého alebo málo zmeneného obsahu. Obsahuje odkaz na predchádzajúci, úplne alebo čiastočne duplicitný záznam z archívu.

```
WARC/1.0
WARC-Type: revisit
WARC-Target-URI: http://consilium.europa.eu/
WARC-Date: 2016-08-30T12:26:15Z
WARC-Payload-Digest: sha1:ENEXF2C2JV62CLQJ5CWYKX2FLFSWSVHG
WARC-IP-Address: 91.194.202.11
WARC-Profile: http://netpreserve.org/warc/1.0/revisit/identical-payload-digest
WARC-Truncated: length
WARC-Refers-To: <urn:uuid:d18647cb-532d-4612-9d3a-47c3fae9d7fc>
WARC-Refers-To-Target-URI: http://consilium.europa.eu/
WARC-Refers-To-Date: 2016-07-01T13:05:18Z
WARC-Record-ID: <urn:uuid:273de3fe-7969-4e01-94f0-bfced099ed40>
Content-Type: application/http; msgtype=response
Content-Length: 310

HTTP/1.1 301 Moved Permanently
Content-Type: text/html; charset=UTF-8
Location: http://www.consilium.europa.eu/
Server: Microsoft-IIS/8.5
X-Powered-By: ASP.NET
Access-Control-Allow-Origin: http://register.consilium.europa.eu
Date: Tue, 30 Aug 2016 12:26:15 GMT
Connection: close
Content-Length: 154
```

Obrázok 6: Záznam typu revisit

„conversion“

Záznamy „conversion“ obsahujú alternatívnu verziu obsahu, teda iný záznam, vytvorený ako výsledok archívneho procesu. Používajú sa pre zastavenie transformácie obsahu, ktorá udržuje životaschopnosť obsahu, v opačnom prípade obsah v pôvodnom formáte zmizne. Pôvodný obsah sa transformuje do rentabilnejšieho formátu, s cieľom udržať informácie použiteľné s existujúcimi nástrojmi a zároveň minimalizuje stratu pôvodných informácií. Záznamy môžu byť vytvorené tak, aby odkazovali na konkrétny zdroj záznamu, ktorý obsahuje transformovaný obsah. Každá transformácia by mala viesť ku kompletnému záznamu bez závislosti na pôvodnom zázname. Pre popis transformácie môžu byť použité metadáta.

```
WARC/1.0
WARC-Type: conversion
WARC-Target-URI: http://www.archive.org/images/logoc.jpg
WARC-Date: 2016-09-19T19:00:40Z
WARC-Record-ID: <urn:uuid:16da6da0-bcdc-49c3-927e-57494593dddd>
WARC-Refers-To: <urn:uuid:92283950-ef2f-4d72-b224-f54c6ec90bb0>
WARC-Block-Digest: sha1:XQMRY75YY42ZWC6JAT6KNXKD37F7MOEK
Content-Type: image/neoimg
Content-Length: 934
```

Obrázok 7: Záznam typu conversion

„continuation“

Navzájom prepojené záznamy na zodpovedajúci obsah z iných súborov WARC. Vytvárajú logicky kompletný záznam pri prekročení limitu súboru WARC. Používajú sa na pokračovanie záznamov rozdelených do viacerých segmentov. Záznam obsahuje pôvodné ID, ktoré definuje začiatok záznamu a posledný „continuation“ záznam musí obsahovať pole „WARC-Segment-Total-Length“. Prvý súbor WARC bude obsahovať prvý segment – záznam typu response.

```
WARC/1.0
WARC-Type: response
WARC-Target-URI: http://www.archive.org/images/logoc.jpg
WARC-Date: 2006-09-19T17:20:24Z
WARC-Block-Digest: sha1:2ASS7ZUZY6ND6CCHXETFVJDENAWF7KQ2
WARC-Payload-Digest: sha1:CCHXETFVJD2MUZY6ND6SS7ZENMWF7KQ2
WARC-IP-Address: 207.241.233.58
WARC-Record-ID: <urn:uuid:39509228-ae2f-11b2-763a-aa4c6ec90bb0>
WARC-Segment-Number: 1
Content-Type: application/http;msgtype=response
Content-Length: 1600

HTTP/1.1 200 OK
Date: Tue, 19 Sep 2006 17:18:40 GMT
Server: Apache/2.0.54 (Ubuntu)
Last-Modified: Mon, 16 Jun 2003 22:28:51 GMT
ETag: "3e45-67e-2ed02ec0"
Accept-Ranges: bytes
Content-Length: 1662
Connection: close
Content-Type: image/jpeg
```

Obrázok 8: Warc typ response začiatok záznamu pri rozdelení

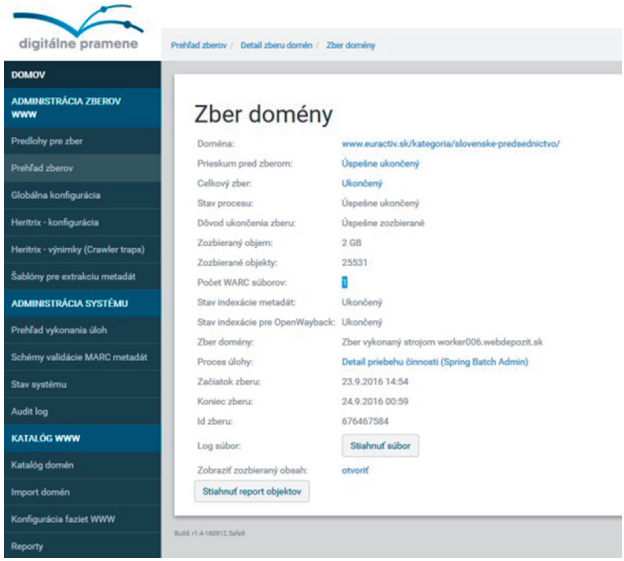
Budúci súbor WARC bude obsahovať pokračovanie záznamu, s poľami pre určenie začiatku segmentu (WARC-Segment-Origin-ID s pôvodným ID), číslo segmentu pre určenie poradia a v prípade posledného segmentu celkovú veľkosť záznamu (WARC-Segment-Total-Length).

```
WARC/1.0
WARC-Type: continuation
WARC-Target-URI: http://www.archive.org/images/logoc.jpg
WARC-Date: 2006-09-19T17:20:24Z
WARC-Block-Digest: sha1:T7HXETFVA92MSS7ZENMFZY6ND6WF7KB7
WARC-Record-ID: <urn:uuid:70653950-a77f-b212-e434-7a7c6ec909ef>
WARC-Segment-Origin-ID: <urn:uuid:39509228-ae2f-11b2-763a-aa4c6ec90bb0>
WARC-Segment-Number: 2
WARC-Segment-Total-Length: 1902
WARC-Identified-Payload-Type: image/jpeg
Content-Length: 302
```

Obrázok 9: Záznam typu continuation

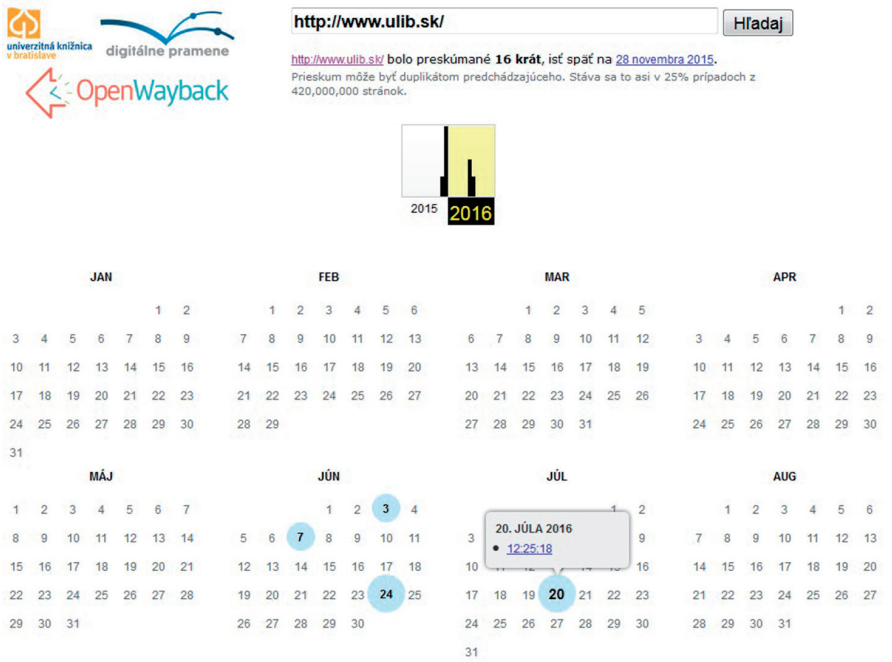
Slovenský webový archív – Webdepozit

Oddelenie DDP (Depozit Digitálnych Prameňov) využíva na zber a archiváciu vybraných webových stránok v podobe súborov WARC voľne dostupné softvérové riešenie Heritrix [7]. Pre účely archívu sú použité SATA disky v objeme 800TB, s predpokladom postačujúceho miesta na minimálne 5 rokov. Modul Web Curator Tool umožňuje cez webové užívateľské rozhranie konfigurovať parametre každého zberu. Heritrix následne zozbiera webový obsah z domén na základe pravidiel zadefinovaných v konfigurácii. Medzi jeho hlavné konfiguračné parametre pre každý webový zber patrí čas, veľkosť a počet dotazov na doménu. Webový obsah sa ukladá ako WARC súbor, pričom modul Deduplikátor kontroluje obsah zozbieraného obsahu a neukladá duplicitný obsah, ktorý už bol zozbieraný a uložený. Počas zberu sa zbierajú aj metadáta z webových stránok a ukladajú sa do katalógu s previazaním na WARC súbory. Na zobrazenie archívneho obsahu webu sa používa open source nástroj OpenWayback [8]. Zbierajú sa primárne html, php, css, js a image formáty. Pri zbere sa rešpektujú nastavenia v súbore robots.txt na strane servera. Pre každú webovú adresu URL zo zberu sa vytvoria vlastné WARC súbory. Jeden WARC má maximálnu veľkosť 2 GB a pri prekročení limitu sa pre danú doménu vytvorí viacero WARC súborov. Archivácia pracovných a log súborov sa komprimuje do súboru ZIP. Vznikne finálny WARC súbor s príponou súboru „warc.gz“.



Obrázok 10: Modul Web Curator Tool

Pri prezeraní archivovaného obsahu zadá používateľ do webového prostredia OpenWayBacku pôvodnú adresu URL stránky, kľúčové slovo alebo vyplní kritéria pre metadáta. Aplikácia zobrazí zoznam archivovaných snímok s časovým rozlíšením, kde používateľ vyberie jednu z nich pre zobrazenie archivovanej stránky.



Obrázok 11: Webové prostredie OpenWayback

Finálnym výsledkom archivácie je archivačný balík SIP (submission information package). SIP balíček je balík dát a metadát akceptovateľný pre LTP (Long Term Preservation) systém. Systém sám vyhľadá vytvorené WARC súbory, ktoré zabalí do SIP balíkov a uloží do cieľového adresára archivácie, kde si ich prevezme Centrálny dátový archív (CDA). CDA je nezávislý archív, spĺňajúci certifikáciu dôveryhodných digitálnych úložísk a informačnej bezpečnosti. Súčasťou SIP balíka je popisný súbor mets-md.xml, v ktorom sa nachádza zoznam všetkých priložených súborov a metadát. Pre archiváciu zozbieraného webového obsahu WARC súborov sa do mets-md.xml vkladajú popisy súborov do súborových tried, ktoré majú v identifikátore uložený názov domény v percentuálnom enkódovaní. Každá takáto skupina obsahuje len súbory patriace danej doméne, čo neskôr uľahčí vyberanie historických záznamov pre konkrétnu stránku. Nie je potrebné sťahovať celé SIP balíky, ale len konkrétny

WARC. Z hľadiska bezpečnosti archívu nemá prístup k vloženým balíkom žiadna iná inštitúcia.

```
crawl name: basic
crawl status: Finished
duration: 2h44m22s978ms

seeds crawled: 2
seeds uncrawled: 0

hosts visited: 50

URIs processed: 6077
URI successes: 5892
URI failures: 0
URI disregards: 185

novel URIs: 3333
duplicate-by-hash URIs: 2559

total crawled bytes: 4236491155 (3.9 GiB)
novel crawled bytes: 985031070 (939 MiB)
duplicate-by-hash crawled bytes: 3251460085 (3.0 GiB)

URIs/sec: 0.6
KB/sec: 419
```

Obrázok 12: Základný popis po skončení zberu

Štatistika súborov WARC vo Webdepozite

Štatistické údaje o počte Warc súborov vo Webdepozite UKB sa sledujú od januára 2016. K 3.10.2016 obsahuje archív Webdepozitu 1726 Warc súborov. Priemerná veľkosť jedného nekomprimovaného Warc je približne 348,5 MB. Ich celková komprimovaná veľkosť je 304 GB a nekomprimovaná 587 GB. Komprimácia teda ušetrí približne 48 % miesta z celkovej veľkosti WARC súborov. Viac informácií je uvedených na stránke www.webdepozit.sk.

Záver

V blízkej budúcnosti plánuje DDP dobudovať modul informačného systému pre extrakciu Open Graph objektov [9] z archívnych balíkov WARC. Projekt je zadaný na vypracovanie ako Diplomová práca na Slovenskej technickej Univerzite v Bratislave, odbor automatizácie a informatizácie v priemysle. Cieľom bude vytvorenie webovej aplikácie pre extrakciu Open Graph objektov z archívnych balíkov a ich transformáciu na databázové objekty archívu. V rámci práce sa bude riešiť vytvorenie autonómnej aplikácie, ktorá bude čítať archívne balíčky WARC a extrahovať z nich Open Graph objekty v zmysle Open Graph protokolu. V aplikácii sa vytvoria konfigurovateľné profily pre metadáta, ktoré budú mapovať Open Graph objekty na objekty v databáze. Aplikácia bude navrhnutá tak, aby ju bolo možné integrovať do existujúceho riešenia DDP ako samostatný modul. Metadáta budú prideľované podľa pravidiel RDA (upravených pre účely projektu Digitálne pramene – webharvesting a archivácia e-Born obsahu) a knihovníckeho formátu MARC 21. Metadáta tvoria hlavnú štruktúru, popis a následnú katalogizáciu pre bibliografické jednotky. Bohatý katalóg s presnou identifikáciou objektov umožní ďalej rozvíjať výskum v archíve Digitálnych prameňov, väčšiu dohľadateľnosť entít v aktuálnom, zmenenom alebo zaniknutom elektronickom obsahu.

Zoznam bibliografických odkazov

[1] ISO 28500: 2009 Information and documentation WARC file format

[2] Sigurðsson, Kristinn. The WARC Format 1.1 [online]. *Blogger*. 17 august 2015, [cit. 2016-09-26]. Dostupné na internete: https://kris-sigur.blogspot.sk/2015/08/the-warc-format-11.html

[3] Github: *IIPC WARC specifications* [online]. [cit. 2016-09-26]. Dostupné na internete: https://github.com/iipc/warc-specifications/

[4] Burner, Mike and Kahle, Brewster. *ARC File Format* [online]. Internet Archive [cit. 2016-09-27]. Dostupné na internete: http://www.archive.org/web/researcher/ArcFileFormat.php

[5] *HTML Unicode UTF-8 Reference* [online]. [cit. 2016-09-27]. Dostupné na internete: http://www.w3schools.com/charsets/ref_html_utf8.asp

[6] *HTTP/1.1 RFC2616* [online]. The Internet Society 1999 [cit. 2016-09-30]. Dostupné na internete: http://www.ietf.org/rfc/rfc2616.txt

[7] *Heritrix* [online]. Heritrix archival crawler project [cit. 2016-09-30]. Dostupné na internete: https://webarchive.jira.com/wiki/display/Heritrix

[8] *OpenWayback* [online]. IIPC [cit. 2016-09-28]. Dostupné na internete: http://netpreserve.org/openwayback

[9] *The Open Graph protocol* [online]. Open Web Foundation Agreement, Version 0.9 [cit. 30 septembra 2016]. Dostupné na internete: http://ogp.me/

WARC 1.1 je skoro tady – co přinese nová verze?

Jaroslav Kvasnica, Národní knihovna ČR

Abstrakt:
WARC je archivní kontejnerový formát, který je základním prvkem webových archivů po celém světě. Po jeho téměř šestileté existenci přichází konsorcium IIPC s jeho minoritní aktualizací na verzi 1.1, která přináší drobné úpravy a opravy chyb verze předchozí. V článku jsou nejen popsány a vysvětleny všechny tyto novinky včetně příkladů a kontextu, ale také je zde popsán samotný proces aktualizace standardu.

1. Úvod

WARC je archivní formát, který je využíván webovými archivy po celém světě. Mnoho nástrojů, ať už jsou vyvíjeny komunitou, nebo jsou komerční, je s formátem WARC spjato. První verze kontejnerového formátu WARC byla oficiálně vydána jako standard ISO v roce 2009. WARC vznikl jako evoluce staršího archivního formátu ARC, který byl vytvořen jako reakce na potřebu webových archivů, které neměly efektivní způsob jak jednoduše uložit svá data stažená z webových stránek, která chtěly archivovat.

Proto je ARC velmi jednoduchý souborový formát, který vznikl za účelem ulehčit manipulaci s velkým objemem souborů a jejich agregováním. V ARCu se nachází jen velmi málo metadat a obsah kontejneru se skládá ze dvou částí. Z hlavičky, která obsahuje jedinečné jméno souboru, IP adresu, čas pořízení souboru, typ a přesnou velikost, přičemž obzvláště důležitou částí hlavičky je také formát záhlaví následujících záznamů, a z těla, které obsahuje URL záznam, záhlaví a samotná data, případně další metadata.[1] To znamená, že každý jednotlivý soubor má přidělenou hlavičku s metadaty.

1 ZACH, Michael. Celosvětový Archiv Internetu a jeho role v získávání, uchování a zpřístupňování webových zdrojů. Praha, 2007. Bakalářská práce. Univerzita Karlova v Praze. Vedoucí práce PhDr. Eva Bratková.

WARC přináší oproti staršímu formátu ARC řadu vylepšení, zejména pokud se jedná o metadatový popis jednotlivých souborů v něm uložených. Formát WARC je jednoduché zřetězení jednoho nebo více tzv. WARC záznamů, které jsou dvojího typu. Prvním typem jsou záznamy pro obsah, tzn. webové stránky, vložené obrázky, informace o přesměrování URL, samostatné soubory atd. Druhým typem jsou záznamy se syntetizovanými daty (metadata), které poskytují další informace o archivovaném obsahu.[2] Každý jednotlivý soubor uložený v kontejneru je opatřený informacemi obsahujícími jeho technický popis a také cestu, kudy pokračovat při zpětné rekonstrukci webové stránky.

Jelikož je formát WARC ISO standardem, tak musí projít procesem, kterým procházejí veškeré standardy při své aktualizaci, tj. hlasováním národních normalizačních organizací, které nejprve musí odhlasovat vůbec možnost aktualizace stávající normy, a pak teprve odhlasovat nové znění normy.

Obecně by měly být všechny standardy revidovány po pěti letech jejich platnosti. V roce 2014 členové ISO organizace hlasovali pozitivně pro revizi standardu WARC. Toto hlasování bylo iniciováno ze strany konsorcia IIPC, které si revizi WARC normy vzalo na starosti. Oficiálně revize, po úspěšném hlasování, začala v roce 2015, kdy se sešly pracovní skupiny, a poprvé došlo k diskuzi, ve které byly probírány problémy, které je potřeba řešit.

Na nové verzi normy se podílely dvě pracovní skupiny, jedna byla ustanovena uvnitř konsorcia IIPC a druhá pocházela přímo z organizace ISO. Nutno dodat, že mnozí členové pracovní skupiny ISO jsou zároveň členy konsorcia IIPC. Koordinátorem pracovních skupiny byl zvolen Clément Oury z Národní francouzské knihovny, který byl i u vzniku první verze formátu WARC.

V současné době je nová verze formátu WARC v etapě, kdy hlasují národní zástupci ISO organizace o jejím přijetí. Toto hlasování obvykle trvá několik měsíců a končí 24. 11. 2016. V případě 100% pozitivního výsledku bude oficiálně vydána nová verze standardu WARC (po finálních redakčních úpravách). Pokud by nějaký člen byl proti jejímu přijetí, bude se konat ještě další kolo připomínkování návrhu.

Již při prvním setkání se pracovní skupiny dohodly, že nová verze WARCu bude pouze minoritní, tzn. že WARC 1.1 nebude obsahovat žádné zásadní koncepční změny. Finální návrh v současné době obsahuje pouze dvě změny, které by se daly označit za novou

2 Web Archiving: Issues and Methods. MASANÈS, Julien. *Web archiving*. New York: Springer, c2006, s. 2-53. ISBN 3540233385-.

funkcionalitu, a pak už se jedná pouze o opravy, upřesnění původních znění nebo odstranění redundantních částí.

2. Rozšíření normy

Asi nejzásadnějšími změnami v normě jsou její rozšíření, která přidávají chybějící funkcionalitu ve starší verzi. V aktuální nové verzi se jedná pouze o dvě taková rozšíření.

2.1. Nové názvové pole pro reduplikaci

Prvním z rozšíření a větších změn, které přináší nová verze formátu WARC, jsou nové názvové pole pro deduplikaci. Každý webový archiv, který pravidelně archivuje webový obsah, se potýká s větší či menší mírou duplicitního obsahu. V praxi se ukazuje, že některé části webových stránek se příliš často nemění, většinou se jedná o statické části, např. loga, patičky, ale i nejrůznější fotogalerie apod. V takových případech se při každé nově vytvořené archivní verzi stahuje opakovaně totožný obsah, a tím samozřejmě dochází k vyšším nárokům na úložné kapacity.

Deduplikace je funkce, která webovým archivům umožňuje, aby totožný obsah nemusela několikanásobně ukládat. Principiálně deduplikace funguje tak, že pokud je na stejné URI stejný obsah jako u předchozí verze, tak se již nestahuje, ale pouze se vytvoří odkaz na již stažený obsah. Tato funkcionalita přináší webovým archivům značné úspory v nárocích na úložné kapacity (v řádu desítek procent, vždy ale záleží na konkrétním zaměření archivu).

Přestože je již v dnešní deduplikace běžně využívána řadou webových archivů po celém světě a aktuální verze standardu WARC tuto funkcionalitu zná, tak její definice není dostatečná. V současné verzi k tomu účelu slouží hlavička “revisit”, která ale umožňuje odkazovat na další záznamy pouze pomocí URI. To s sebou nese problém, že odkazovaný obsah musí mít vždy stejné URI jako obsah samotný a zároveň se musí spolehnout na to, že na odkazovaném URI je nejnovější verze původního obsahu.

Zjednodušeně řečeno WARC 1.0 umožňuje pouze deduplikaci na úrovni URL, tzn. že pokud na stejné URL je stejný obsah jako u předchozích archivních verzí, pak je možné obsah nahradit odkazem na předchozí verze. U nové verze je možné deduplikovat

stejný obsah na různých URL, např. stejný obrázek na různých webových stránkách tzv. prostorová deduplikaci, která je je možná díky kombinaci URI a časového údaje.

Nová verze přináší tato nová názvová pole, která řeší problém s prostorovou deduplikací:[3]

`WARC-Refers-To-Target-URI`

V tomto poli je zapsáno URI záznamu, který je deduplikován. Pole by mělo být využíváno pouze pro hlavičku typu "revisit".

`WARC-Refers-To-Date`

V tomto poli by měl být zapsán časový údaj deduplikovaného záznamu. Pole by také mělo být využíváno pouze pro hlavičku typu "revisit".

Díky této nové funkcionalitě je možné velmi zefektivnit deduplikaci, webové archivy, tak mohou ušetřit ještě více úložného prostoru a tím celý proces archivace výrazně zlevnit.

2.2. Rozšíření možnosti zápisu časových údajů

Druhou z nových funkcí, kterou nová verze přináší, je větší variabilita pro zápis časového údaje v elementu WARC-Date. Současná verze umožňovala pouze jeden zápis, který byl striktně definován jako řetězec formátový jako "`YYYY-MM-DDT:hh:mm:ssZ`"[4]. Tato definice neumožňovala žádnou variabilitu a u neúplných časových údajů musel být zápis doplňován nulami.

V nové verzi autoři přistoupili k začlenění normy ISO8601[5] a zároveň umožnili variabilní počet znaků řetězce. To znamená, že nově je možné WARC-Date specifikovat na jakékoliv úrovni granularity popsané ve výše uvedené časové normě.

3 Proposal for Standardizing the Recording of Arbitrary Duplicates in WARC Files 1.0. INTERNATIONAL INTERNET PRESERVATION CONSORTIUM. WARC specifications [online]. 2016 [cit. 2016-10-17]. Dostupné z: https://iipc.github.io/warc-specifications/specifications/warc-deduplication/recording-arbitrary-duplicates-1.0/

4 ISO 28500:2009: Information and documentation – WARC file format. 1. edition. Londýn, 2009.

5 ISO 8601 je mezinárodní standard pro zápis data a času. Tuto normu využívá i W3C pro webové standardy.

Příklad jediného možného zápisu ve verzi 1.0:

```
2007-11-02T15:20:44Z
```

Příklady možností zápisu ve verzi 1.1:

```
2007-11-02T15:20:44Z
2007-11
2007-11-02T15:20:44.5Z
2007-11-02T15:20:44.23453Z
```

3. Oprava chyby – zápis URI v hlavičce

Stávající standard sám sobě odporoval v zápisu URI v hlavičkách formátu WARC. V samotné definici standard nařizoval použití závorek "<" a ">" při zápisu URI. Viz následující příklad:

```
uri      = „<" <'URI' per RFC3986> ">"
WARC-Target-URI: <http://www.archive.org/images/logoc.jpg>
```

Nicméně v příkladech bylo URI uvedeno bez těchto závorek. Viz následující příklad:

```
WARC-Target-URI: http://www.archive.org/images/logoc.jpg
```

V praxi se ujal druhý typ zápisu, který byl uvedený v příkladech i veškeré nástroje mají implementovaný tento styl zápisu. Proto bude v nové verzi standardu použit styl zápisu bez závorek.

4. Odstranění redundantních částí

4.1. Definice MIME typu pro WARC

Stávající standard pro WARC 1.0 obsahuje definici MIME typu[6] pro samotný formát WARC (application/warc, application/warc-fields). Tato definice bude z nové verze

6 MIME typ je dvoudílný identifikátor formátu souborů na Internetu.

standardu vyjmuta, protože není běžným zvykem, aby definice MIME typu byla součástí standardu formátu. IIPC v budoucnu navrhne, aby byl WARC MIME typ zařazen do registru MIME typů, který spravuje IANA (The Internet Assigned Numbers Authority).

4.2. Názvová konvence pro soubory WARC

Další odstraněná část se dotýká názvové konvence pro soubory WARC. Ve verzi 1.0 byla zmínka, o tom, že členové konsorcia IIPC by měly používat u svých souborů prefix 'iipc_'. Toto nařízení se ovšem v praxi neujalo a tak v nové verzi bude již vypuštěno.

5. Úpravy stávajícího znění standardu

5.1. Vytváření nových názvových polí

Již verze 1.0 umožňuje přidávat, v případě potřeby webového archivu, vlastní názvové oblasti do hlaviček. Nová verze k tomuto přidává doporučení konzultace s konsorciem IIPC, tak aby nedocházelo např. k vytváření nových názvových polí, které již vytvořil jiný webový archiv a předešlo se tak kolizím v názvech a jednotlivé WARC soubory zůstaly vzájemně kompatibilní.

5.2. Zaměření standardu, bezpečnostní otázky a hlavička "warcinfo"

Poslední část změn se věnuje drobným úpravám znění standardu. V části o zaměření standardu je nyní reflektováno, že se po dobu své existence WARC rozšířil i mimo webové archivy. U části o bezpečnosti je zdůrazněno, že konvence pro zaznamenání informací k HTTPS protokolu nejsou otázkou tohoto standardu.

A v neposlední řadě je umožněno nyní v hlavičce "warcinfo" zapsat použitý algoritmus u kontrolního součtu. Viz příklad:

```
WARC-Block-Digest: sha1:AB2CD3EF4GH5IJ6KL7 MN8OPQ

WARC-Block-Digest: sha1_Base32:AB2CD3EF4GH5IJ6KL7 MN8OPQ
```

6. Závěr

Závěrem je nutno dodat, že z výčtu výše uvedených změn je evidentní, že se jedná pouze o minoritní aktualizaci standardu. Nicméně jde o úpravy, které reflektují 6 let používání souborové formátu desítkami webových archivů po celém světě, a proto rozsah této aktualizace neubírá na její důležitosti. Zároveň lze s jistotu říct, že veškeré novinky a změny budou komunitou zapracovány do všech relevantních nástrojů (zejména jde-li o sklízeč Heritrix a zobrazovací aplikace OpenWayback), tudíž nebude žádný problém pro webové archivy na nový WARC 1.1 přejít.

Použitá literatura

Proposal for Standardizing the Recording of Arbitrary Duplicates in WARC Files 1.0. INTERNATIONAL INTERNET PRESERVATION CONSORTIUM. WARC specifications [online]. 2016 [cit. 2016-10-17]. Dostupné z: https://iipc.github.io/warc-specifications/specifications/warc-deduplication/recording-arbitrary-duplicates-1.0/

ISO 28500:2009: Information and documentation – WARC file format. 1. edition. Londýn, 2009.

Web Archiving: Issues and Methods. MASANÈS, Julien. *Web archiving*. New York: Springer, c2006, s. 2-53. ISBN 3540233385-.

ZACH, Michael. Celosvětový Archiv Internetu a jeho role v získávání, uchování a zpřístupňování webových zdrojů. Praha, 2007. Bakalářská práce. Univerzita Karlova v Praze. Vedoucí práce PhDr. Eva Bratková.

Zoznam autorov

Ing. Milan Rakús
Univerzitná knižnica v Bratislave

Mgr. Bibiána Žigová
Univerzitná knižnica v Bratislave

Mgr. Jan Hutař, Ph.D.
Archives New Zealand

PhDr. Ladislav Cubr
Národní knihovna České republiky

Bakk. techn. Peter Bubestinger
Osterreichische Mediathek

Bc. Andrej Bizík
Univerzitná knižnica v Bratislave

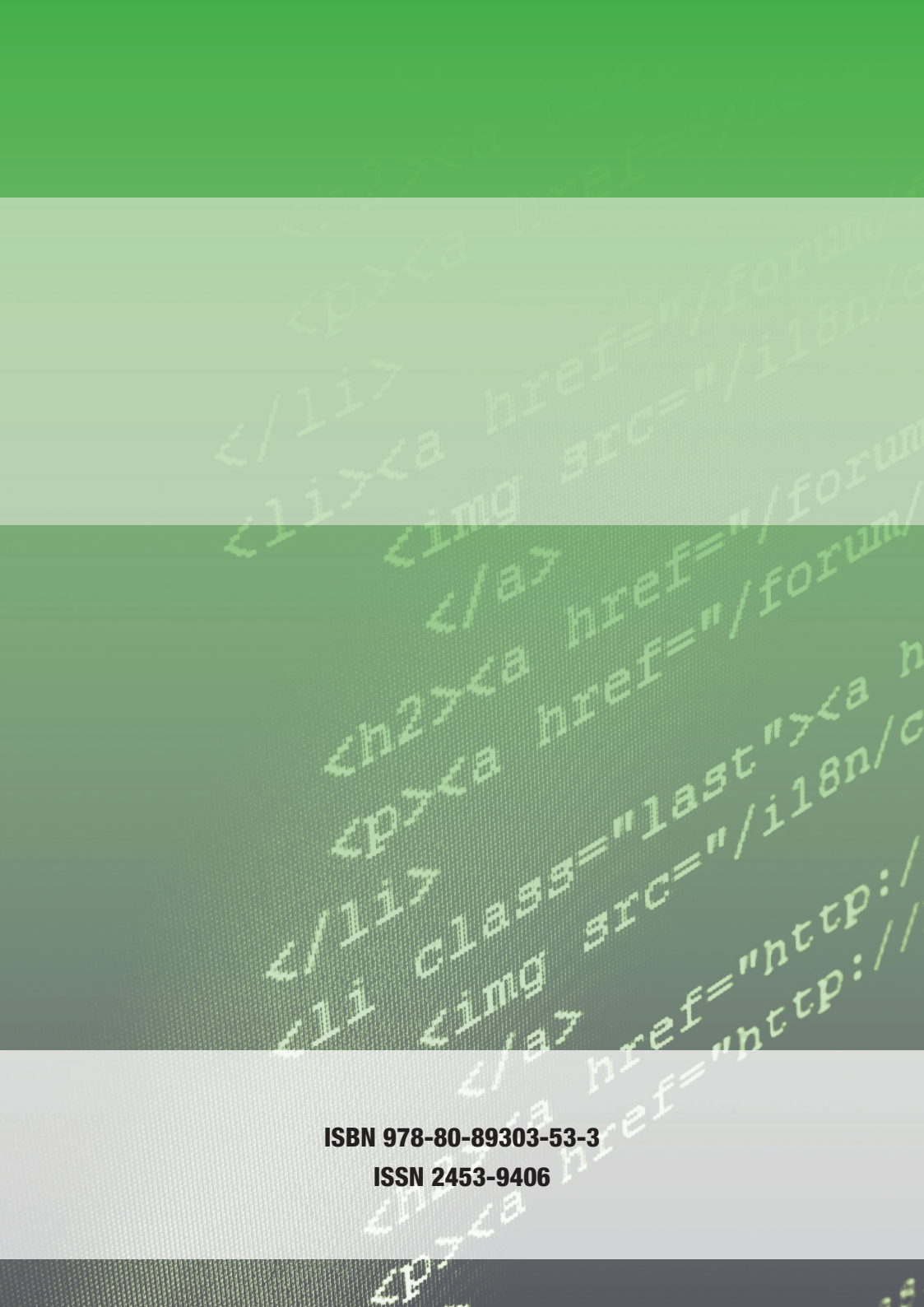
Mgr. Jaroslav Kvasnica
Národní knihovna České republiky

CDA 2016

Formátové výzvy LTP

Vydala Univerzitná knižnica v Bratislave
Prvé vydanie. Počet strán 102.
Sadzba: DOLIS, s.r.o., Bratislava
Tlač: DOLIS, s.r.o., Bratislava

ISBN 978-80-89303-53-3
ISSN 2453-9406



ISBN 978-80-89303-53-3
ISSN 2453-9406