MOŽNOSTI KONVERZÍ PDF SOUBORŮ PRO POTŘEBY DLOUHODOBÉ ARCHIVACE ANEB S PDF/A NA VĚČNÉ ČASY?

Mgr. Jana Hrzinová; jana.hrzinova@nkp.cz; (Národní knihovna České republiky)
Bc. Václav Jiroušek; vaclav.jirousek@nkp.cz; (Národní knihovna České republiky)

V rámci přípravy na příjem povinného (či dobrovolného) výtisku elektronických publikací realizovala Národní knihovna ČR v roce 2022 na vzorku dat pilotní studii zaměřenou mimo jiné na možnost konverzí souborů PDF do archivačních verzí PDF/A. Závěry byly publikovány v článku „Výběr archivačních formátů pro povinný depozit e-publikací v ČR: EPUB a PDF/A jako řešení?". Otázka, jak naložit s neúspěšně převedenými soubory, se stala základem pro navazující analýzu realizovanou v roce 2023, jejíž výstupy jsou předmětem tohoto článku. Na základě detailnějšího rozboru a popsání jednotlivých kategorií chyb detekovaných validačními nástroji po prvotním převodu a vyzkoušení dalších možností nastavení konverzních nástrojů se cílem této navazující analýzy stalo zmapování možností a postupů potřebných k zachování stanovených signifikantních vlastností souborů, převáděných do formátů vhodných pro dlouhodobou archivaci.

http://doi.org/10.52036/1335793X.2023.3-4.30-36

ÚVOD

Předkládaný text navazuje na článek autorů publikovaný v IT*lib* 1–2/2022 pod názvem *„Výběr archivačních formátů pro povinný depozit e-publikací v ČR: EPUB a PDF/A jako řešení?".* Úvodní článek zkoumal zastoupení jednotlivých souborových formátů na českém trhu prostřednictvím přidělených identifikátorů ISBN, vhodnost jednotlivých formátů pro dlouhodobou archivaci a představoval výsledky pilotní analýzy zaměřené na převoditelnost PDF souborů do jejich archivní varianty PDF/A. V navazující studii se podrobněji zabýváme kategoriemi chyb detekovaných při konverzích PDF na vzorku dat spravovaných Národní knihovnou ČR do PDF/A a možnostmi jejich řešení. Posuzovány jsou možnosti konverzí do jednotlivých variant formátu PDF/A i jejich vhodnost pro dlouhodobou archivaci a zpřístupnění. S ohledem na stále chybějící legislativu, která by v České republice zaváděla povinnost odevzdávat elektronické publikace do knihoven k jejich dlouhodobému uchovávání, připravuje Národní knihovna ČR v první fázi systém odevzdávání založený na dobrovolnosti. Vzhledem k minimálnímu užívání formátu PDF/A při distribuci elektronických dokumentů na knižním trhu považujeme schopnost příjmu rozmanitých variant formátu PDF a jejich normalizace do archivních souborů PDF/A za nezbytnou pro maximální účinnost celého systému, tedy zachycení a uchování co největší části písemného kulturního dědictví šířeného v elektronické podobě.

VERZE FORMÁTU PDF/A A JEJICH SPECIFIKA

PDF/A

PDF/A je otevřený souborový formát, určený pro dlouhodobou archivaci digitálních dokumentů. Definuje vlastnosti PDF tak, aby byly minimalizovány veškeré potenciální hrozby budoucí zobrazitelnosti dat. Jedná se tedy o zúžený formát PDF, zabezpečující zachování všech informací bez závislosti na platformě. Jedná se o samostatné, samopopisné soubory, obsahující veškeré potřebné informace pro jejich zobrazení bez ohledu na použité technologie. Soubory nesmí být zejména šifrované, chráněné heslem a obsah musí být plně přístupný. Nejsou povolené například vložené multimediální a jiné prvky, které by vyžadovaly další software pro jejich spuštění. Jsou kladeny vyšší nároky na všechny informace, které dokument obsahuje, a proto musí být vložené všechny fonty a popsané glyfy, dále se používají ICC profily a definované metadatové formáty. Obsažená metadata jsou důležitá i z důvodu správné identifikace verze profilu PDF/A, kterou bez metadat není možné jednoduše určit. Hodí se zejména pro statické dokumenty, u kterých je důležité zachování rozvržení, vzhledu, textového obsahu apod. Naopak není vhodný pro dokumenty, které obsahují multimediální prvky. Formát jako takový zajišťuje autenticitu dokumentu tím, že obsahuje všechny informace jako při jeho vzniku a zamezuje budoucím

modifikacím. Všechny přednosti toho archivačního formátu můžeme do jisté míry brát také jako jeho nevýhodu. Skutečnost, že některé funkce jsou zakázané (audio, video, JavaScript apod.), podobně jako požadavek na vložení fontů, mohou být značně omezující jak při prvotní tvorbě elektronických dokumentů, tak zejména při pozdějších konverzích dříve vytvořených souborů PDF, které je potřeba do archivační podoby převádět. Požadavky na vložené prvky a samopopisnost souborů mohou zároveň výrazně ovlivňovat výsledné velikosti souborů určených pro dlouhodobou archivaci (Fanning 2017; Oettler 2013; New York State Archives 2013).

Vedle první verze PDF/A-1 (ISO 19005-1:2005) vznikly v návaznosti na revize mateřského formátu PDF a rovněž z důvodu potřeby rozšíření standardu o nové funkce do současnosti tři další základní verze formátu: PDF/A-2 (ISO 19005-2:2011), PDF/A-3 (ISO 19005-3:2012) a nejnovější PDF/A-4 (ISO 19005-4:2020). V rámci verzí pak existuje ještě několik úrovní souladu s normou, které dále přiblížíme.

PDF/A-1 (ISO 19005-1:2005)

PDF/A-1 vychází z funkcionalit verze PDF 1.4. Tento formát vyžaduje vložené písmo, vložené barevné prostory, metadata ve formátu XMP, povinné mapování textu na standard Unicode, přesně popsanou logickou strukturu dokumentu, a naopak nepovoluje šifrování, LZW kompresi (i z důvodu, že tento kompresní algoritmus byl až do roku 2004 zatížen patentem), vložené soubory, odkazy na externí obsah, průhlednost, multimédia nebo JavaScript.

PDF/A-2 (ISO 19005-2:2011)

PDF/A-2 je založené na verzi PDF 1.7 (ISO 32000-1:2008). V základu vyžaduje všechna pravidla předchozí verze, ale formát je rozšířený o některé další funkcionality. Mimo podpory vložených fontů typu PostScript podporuje nově také fonty OpenType. Dále podporuje kompresi obrazů JPEG2000, transparentnost a použití více vrstev a stává se také kontejnerovým formátem – dovoluje vkládání dalších souborů ve formátu PDF/A.

PDF/A-3 (ISO 19005-3:2012)

Tento formát vychází stejně jako PDF/A-2 z verze PDF 1.7 (ISO 32000-1:2008), předchozí verzi nicméně rozšiřuje o některé nové funkce. Zatímco ve verzi PDF/A-2 mohou být vloženy pouze soubory ve formátu PDF/A, verze PDF/A-3 dovoluje i jiné formáty vložených souborů. Díky tomu mohou soubory obsahovat například původní dokumenty, ze kterých byla jejich archivní verze v PDF/A vytvořena. V případě vložení dalšího obsahu formát nicméně garantuje zachování pouze statické vizuální reprezentace vloženého obsahu.

ÚROVNĚ SOULADU S NORMOU U PDF/A-1 AŽ PDF/A-3

Každá z verzí archivního formátu PDF/A-1 až PDF/A-3 má k dispozici několik úrovní, které definují míru souladu s danou normou. Úroveň A (Accessible) vyžaduje splnění všech požadavků daných normou včetně zachování logické struktury dokumentu, definice jazyka a mapování do znakové sady Unicode. Úroveň B (Basic) vyžaduje přesnou vizuální reprodukci, ale informace o logické struktuře nebo sémantické informace nejsou vyžadovány. Tato úroveň nezaručuje, že bude zachován

<table>
<tr><th>verze PDF/A</th><th colspan="2">charakteristika</th></tr>
<tr><td>PDF/A-1a</td><td colspan="2">vložené fonty, specifikace jazyka, hierarchická struktura dokumentu, popisný text pro obrazy, mapování znaků do Unicode</td></tr>
<tr><td>PDF/A-1b</td><td colspan="2">vložené fonty</td></tr>
<tr><td>PDF/A-2a</td><td colspan="2">odpovídá 1a s vylepšeným tagováním</td></tr>
<tr><td>PDF/A-2b</td><td colspan="2">odpovídá 1b a dovoluje použití průhlednosti</td></tr>
<tr><td>PDF/A-2u</td><td colspan="2">odpovídá 2b a vyžaduje mapování do Unicode</td></tr>
<tr><td>PDF/A-3a</td><td>odpovídá 2a</td><td rowspan="3">a povoluje vložené soubory jako součást PDF/A</td></tr>
<tr><td>PDF/A-3b</td><td>odpovídá 2b</td></tr>
<tr><td>PDF/A-3u</td><td>odpovídá 2u</td></tr>
</table>

Tabulka 1 Charakteristiky jednotlivých verzí PDF/A-1 až PDF/A-3

tok textu, a dokument tak může být například méně přístupný pro osoby se zrakovým znevýhodněním. Úroveň U (Unicode) je dostupná pro verze PDF/A-2 a PDF/A-3 a opět nevyžaduje informace o logické struktuře, nicméně text je mapován do znakové sady Unicode (Fanning 2017; Oettler 2013; New York State Archives 2013).

Charakteristiky jednotlivých verzí archivních formátů PDF/A-1 až PDF/A-3, včetně úrovně jejich souladu s příslušnou normou PDF/A, shrnuje tabulka 1 (Klindt 2017).

PŘEVODY DO VERZÍ FORMÁTU PDF/A A NÁSLEDNÁ ANALÝZA CHYB

V loňském roce byla realizována úvodní analýza možností konverzí nad soubory PDF odevzdanými do Národní knihovny ČR v rámci dobrovolného příjmu elektronických publikací. Pro zkušební převody byl zvolen formát PDF/A-1b a účelem konverzí bylo získat představu o kvalitě dodaných souborů a možnostech převodu do archivačního formátu. Úspěšnost úvodní zkušební konverze dosáhla nepříliš uspokojivých 81,3 %, logickým následujícím krokem se proto stala analýza chyb a jejich řešení. Současně došlo nově k ověření úspěšnosti konverzí do dalších verzí archivačního formátu PDF/A.

Původní vzorek e-knih ve formátech PDF, dobrovolně odevzdaných do Národní knihovny ČR, byl v roce 2023 rozšířen o další soubory z provenience Městské knihovny v Praze, Knihovny Akademie věd a Studijní a vědecké knihovny Hradec Králové. V případě Městské knihovny v Praze se nejedná pouze o její produkci, ale také o produkci různých nakladatelů, jako je Paseka, Host, CZNIC apod. Celý zkoumaný vzorek čítal 637 souborů.

HROMADNÝ PŘEVOD DO FORMÁTŮ PDF/A-1A, A-1B, A-2B, A-3B, A-2U

Prvním krokem analýzy byl hromadný převod všech původních PDF souborů do tří základních verzí archivačního formátu PDF/A, oproti loňskému roku tak byly konverze rozšířeny o formáty PDF/A-2b, PDF/A-3b a PDF/A-2u a jako doplnění analýzy byl proveden i převod do striktního formátu PDF/A-1a. Jako nástroj převodu byl použit Adobe Acrobat PRO 2020 ve verzi 2020.005.30524. Pro převod byly použity výchozí profily Adobe Acrobatu pro jednotlivé standardy PDF/A s původním nastavením a možností aplikování přednastavených oprav pro splnění požadavků převodu na jednotlivé archivační formáty. Zabudovaný nástroj Kontrola před výstupem[1] obsahuje předdefinované opravy, týkající se například barev, písem, obrázků apod. v souladu s mezinárodními standardy. V případě převodů do PDF/A-1b, PDF/A-2b a PDF/A-3b obsahuje nástroj dvě varianty převodu. Pro účely analýzy byla nejprve vybrána varianta bez záložního převodu, který by v případě neúspěchu převedl problematická místa na obrazy. Adobe Acrobat umožňuje nastavení vlastních hromadných akcí a definování cílových složek v případě úspěchu nebo neúspěchu dané operace. Převáděný dokument se do složky s neúspěšnými převody dostane i v případě, že převod dopadne s varováním nebo neúspěšnou dílčí opravou, která ovšem striktně nebrání splnění podmínek pro standardizovaný formát. V této složce tak mohou skončit i ty soubory, které jsou validačním nástrojem vyhodnoceny jako validní. Ke každé konverzi byl také vygenerován report obsahující záznam o provedených opravách, případných chybách a informace o daném souboru. Při nastavení generování zprávy o převodu je možné volit mezi formátem PDF, XML a TXT. Po dokončení převodu byla u souborů ověřena validita vůči příslušnému profilu pomocí nástroje veraPDF ve verzi 1.20.1. Výsledky prvotní konverze a validace jsou patrné v tabulce 2.

VYHODNOCENÍ CHYBNÝCH PŘEVODŮ

Z podrobnější analýzy zpráv o převodu, a také samotných výsledných souborů, vyplynulo několik skutečností. Jako nejproblematičtější dopadl podle očekávání

	neúspěšně převedeno	úspěšně převedeno	z toho nevalidní podle veraPDF
PDF/A-1a	517	120	0
PDF/A-1b	147	490	23
PDF/A-2b	139	498	127
PDF/A-3b	128	509	132
PDF/A-2u	145	492	111
PDF/A (libovolné)	**82**	**555**	**28**

Tabulka 2 Výsledky automatických konverzí dle jednotlivých cílových formátů.

převod do archivačního formátu PDF/A-1a, který má striktní požadavky zejména na popsání přesné logické struktury dokumentu. Z celkového počtu se podařilo úspěšně převést pouze 120 souborů (18,8 %), úspěšně zkontrolovaných validačním nástrojem. Neúspěšně převedené soubory nejčastěji neobsahovaly povinné součásti, týkající se právě logické struktury a MarkInfo. Problémy při převodech do ostatních formátů by se potom daly rozdělit do několika skupin.

První skupinou jsou problémy samotných vstupních souborů. Ze zkoumaného vzorku bylo 30 vstupních souborů chráněno heslem a 6 souborů již bylo v archivačním formátu PDF/A. V případě dokumentů chráněných heslem skončí soubor ve složce s neúspěšnými převody, ale protože převod neproběhne, není ani vytvořena zpráva o převodu. Zjištění ochrany heslem je možné buďto z vlastností dokumentu, anebo při pokusu o druhý kontrolní převod, kdy program uživatele vyzve k zadání hesla. Tato skupina problémů ovlivnila samozřejmě neúspěch migrace do formátu PDF/A-1a i ostatních verzí formátu.

Druhou skupinou jsou převody neúspěšné z důvodu nepodporování některé z vlastností vybraným archivačním profilem. Mezi tyto detekované problémy patří použití standardu pro popis vektorových písem OpenType (vyskytlo se u 36 vstupních souborů) a přítomnost vložených souborů (13 vstupních souborů). Fonty OpenType[2] jsou u PDF podporovány od verze 1.6, a nemohou tak být použity ve verzi PDF/A-1. Řešením pro tento problém je použití konverze do verze PDF/A-2 (v našem případě PDF/A-2b).

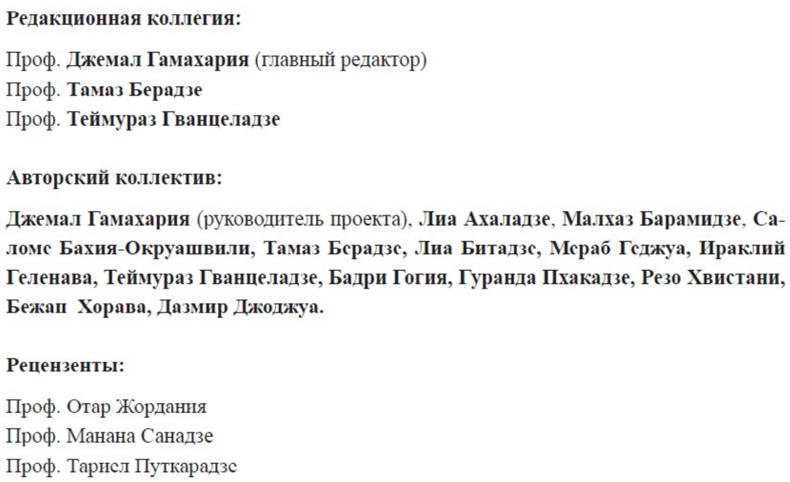
V případě přítomnosti vložených souborů – u analyzovaných dokumentů se jednalo o soubory XML a soubor .joboptions (soubor s parametry pro tvorbu PDF používaný nástrojem Adobe Distiller) – je pro konverzi nutné použití verze PDF/A-3, která jejich přítomnost dovoluje. Přítomnost písma OpenType a vložených souborů je možné ověřit nástrojem Kontrola před výstupem a spuštěním analýzy souboru před samotným převodem.

Další chybou, navázanou na absenci podpory ze strany verzí formátu PDF/A-1 (resp. PDF 1.4), je problém syntaxe, kdy pole obsahuje více než 8191 prvků[3]. Tento problém se vyskytl celkem u 16 souborů a po jejich převodu do PDF/A-2 byly úspěšně zvalidovány. Jako vhodné se jeví provedení dávkové kontroly nad soubory před samotnou konverzí a provedení analýzy výsledných reportů. Při použití formátu XML se dají soubory jednoduše prohledávat a detekovat tak zdrojová PDF, která pro konverzi vyžadují použití formátu PDF/A-2b, respektive PDF/A-3b. Problémy s použitým písmem a glyfy tvoří samostatnou skupinu problémů, které můžeme rozdělit do několika podskupin. Aby byly splněny podmínky standardu, musí být všechna písma vložená a musí být také legálně vložitelná pro neomezené a univerzální vykreslování (Library of Congress 2020). Při analýze chybných převodů se problém s vloženým písmem vyskytl celkem u 28 souborů. Pokud se jednalo o volně dostupné písmo, které nebylo nainstalované v systému, po jeho instalaci byl převod již úspěšný. Takto se podařilo dodatečně zkonvertovat dalších 7 souborů. U zbylých 21 souborů se jednalo o běžně nedostupná písma s licencí, případně písmo zakazovalo vkládání.

Dílčím řešením pro takovéto soubory je konverze se záložním převodem, kdy jsou problematické strany převedeny na obrazy. Je možné zvolit kvalitu komprese a při vypnutí aplikování opravy, která písmo vkládá, je výsledkem validní PDF/A soubor. Nicméně při následné vizuální kontrole takto převedených souborů jsou patrné rozdíly v kvalitě z důvodu přechodu z vektorového zobrazení na rastr. U celkem 8 takto převedených souborů se jednalo o chybu na titulní nebo zadní straně (nejčastěji v logu nakladatele) a obsah tak nebyl ovlivněn, včetně schopnosti vykopírování textu. Je k další diskusi, zda tyto drobné odchylky budeme tolerovat ve prospěch dlouhodobé archivace, nakolik může výsledný dokument nahradit originální PDF a zda je takový soubor PDF/A případně vhodný ke zpřístupnění uživatelům prostřednictvím digitální knihovny, nebo má sloužit důsledně pouze jako záložní archivační kopie. U 12 zbylých dokumentů byly rozdíly v kvalitě zásadní a týkaly se již samotného obsahu dokumentů, u některých došlo ke ztrátě znaků nebo celých částí textu (viz obrázky 1 a 2), negativně ovlivněna byla také možnost vykopírování a prohledatelnosti textu.

Podobným problémem jsou nekonzistentní informace o šířce glyfu (4 výskyty), které se v žádném ze zachycených případů nepodařilo opravit, a východiskem byl pouze převod problematických stran na obrazy. Pro některé výskyty existuje řešení ručního zásahu přímo do zdrojového souboru (viz. Hoppenheit 2018[4]), nicméně v našem případě se tyto chyby nepodařilo ani touto cestou opravit.

V průběhu analýzy se objevily i chyby s ojedinělými výskyty, které se podařilo eliminovat zařazením některé z dostupných oprav nástrojem Kontrola před výstupem. Na chybu „obraz JPEG2000 používá nesprávnou hodnotu položky METH“ (2 výskyty) byla použita oprava „převést obrazy JPEG2000 na JPG“, v případě problémů s barevným prostorem CMYK byl použit převod do prostoru sRGB (2 výskyty).



Редакционная коллегия:

Проф. **Джемал Гамахария** (главный редактор)
Проф. **Тамаз Берадзе**
Проф. **Теймураз Гванцеладзе**

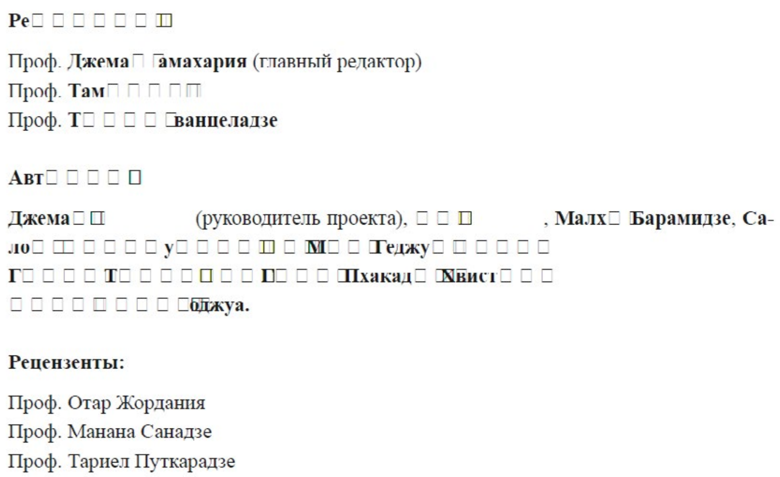
Авторский коллектив:

Джемал Гамахария (руководитель проекта), **Лиа Ахаладзе, Малхаз Барамидзе, Саломе Бахия-Окруашвили, Тамаз Берадзе, Лиа Битадзе, Мераб Геджуа, Ираклий Геленава, Теймураз Гванцеладзе, Бадри Гогия, Гуранда Пхакадзе, Резо Хвистани, Бежан Хорава, Дазмир Джоджуа.**

Рецензенты:

Проф. Отар Жордания
Проф. Манана Санадзе
Проф. Тарнел Путкарадзе

Obrázek 1 Původní soubor PDF před konverzí



Ре□□□□□□□

Проф. **Джема□ □амахария** (главный редактор)
Проф. **Там□□□□□**
Проф. **Т□□□□□ванцеладзе**

Авт□□□□□

Джема□□ (руководитель проекта), □□□ , **Малх□ Барамидзе**, **Са-**
ло□□□□□□□у□□□□□□□М□□□Геджу□□□□□□
Г□□□□Т□□□□□□□□□□□□Пхакад□□Хвист□□□
□□□□□□□□□□оджуа.

Рецензенты:

Проф. Отар Жордания
Проф. Манана Санадзе
Проф. Тарнел Путкарадзе

Obrázek 2 Výsledný soubor po převodu do PDF/A – navzdory využití možnosti nahrazení textu za obrázek došlo ke ztrátě významné části obsahu

Ojedinělá chyba, kdy znak odkazoval na glyf .notdef, byla v rámci analýzy vyřešena převodem písma v původním souboru na obrysy a jeho přeuložením. Následná konverze do PDF/A proběhla bez problémů a vizuální kontrola i vykopírování textu byly v tomto případě v pořádku. Všechny tyto opravy nicméně mohou zasáhnout do autenticity souborů a je opět k následné diskusi, v jakém rozsahu jsou ve prospěch dlouhodobé archivace legitimní a nakolik může výsledný soubor plnohodnotně nahradit originální PDF.

Poslední kategorií jsou chyby vyhodnocené validačním nástrojem veraPDF, které Adobe Acrobat nedetekoval a soubory považoval za úspěšně převedené a validní. U 127 souborů převedených do formátu PDF/A-2b nástroj veraPDF detekoval chybu formátu metadat XMP (Extensible Metadata Platform, standardizovaný způsob pro zápis metadat k souborům), konkrétně porušení pravidla validace 6.6.2.3.2.[5] 89 souborů bylo úspěšně převedeno do formátu PDF/A-1b. 3 soubory se nepodařilo úspěšně převést z důvodu kombinace s jinou chybou, zmíněnou výše. 35 souborů nebylo možné převést do formátu PF/A-1b z důvodu přítomnosti OpenType fontu (verzi PDF/A-1b nepodporovaného), nebo jiného podobného omezení. V rámci experimentu bylo zjištěno, že pokud původní soubor převedeme nejprve s chybou do formátu PDF/A-1b (dojde k jeho částečné opravě a přeuložení) a tento soubor dále převedeme na PDF/A-2b, je převod úspěšný, podle veraPDF již validní a po vizuální kontrole nebyly nalezeny žádné rozdíly.

K uvedené chybě metadat XMP u celé skupiny 127 souborů je nutno dodat, že není detekována ani Adobe Acrobatem, ani žádným z nástrojů zapojených v dodatečně použitém online validátor Národního archivu ČR[6]. Nelze tedy vyloučit, že se jedná o chybu na straně vali-

problém při převodu	počet
chráněno heslem	30
původní soubor PDF/A	6
použito písmo OpenType	36
obsahuje vložené soubory	13
problém syntaxe (pole obsahuje více než 8191 prvků)	16
písmo není vložené/problémy s glyfy	32
obraz JPEG2000 používá nesprávnou hodnotu položky METH	2
problém s barevným prostorem CMYK	2

Tabulka 3 Přehled detekovaných chyb

dátoru veraPDF. Samotná validace XMP může být ostatně poměrně problematická, jelikož specifikace XMP není příliš striktní (Fischer, Lundell, Gamalielsson 2021).

VYHODNOCENÍ

Díky analýze zdrojových souborů a chyb získaných při prvotním převodu jsme byli schopni identifikovat soubory, pro něž je nutné použít jiný formát než PDF/A-1b, použitý pro úvodní analýzu realizovanou v loňském roce. Pokud ze statistiky vyloučíme 30 souborů chráněných heslem, dostaneme 579 úspěšně převedených (včetně 6 již dodaných souborů PDF/A) a validních souborů z celkového počtu 607, tedy 95,4 %. Úspěšně převedené dokumenty byly vizuálně zkontrolovány a byla ověřena schopnost zachování textu po vykopírování. Velké množství původně dodaných PDF mělo textovou vrstvu poměrně nekvalitní, ale i přes tuto skutečnost nebyly při vykopírování zaznamenány rozdíly. Zbylých 28 souborů mělo v drtivé většině problém s písmem a glyfy a podařilo se je převést pouze za pomoci použití záložního převodu na obrazy, v některých případech s malými, ale jinde i významnými vizuálními ztrátami, včetně ztrát znaků a textu, což nelze považovat za úspěšný převod. Ve výsledku tak bylo úspěšně převedeno 90,9 % souborů, přičemž mírnou většinu z neúspěšně převedených (30 z 58) tvoří dokumenty chráněné heslem, které je možné při příjmu poměrně jednoduše detekovat a odmítnout.

V průběhu analýzy se jako největší problém projevily chyby s písmem. Vzhledem k tomu, že většina nástrojů na tvorbu elektronických dokumentů podporuje exporty souborů v archivačním formátu PDF/A, bylo by vhodné, aby soubory určené k archivaci byly vytvářeny právě v tomto formátu, jelikož tvůrce disponuje fonty, a pokud to nezakazuje přímo jejich licence, mohou být do souborů rovnou vkládány. Později roste pravděpodobnost, že archivátor nebude mít fonty k dispozici, a bude se muset spokojit s méně kvalitní převedenou verzí. Vzhledem k tomuto, a také s ohledem na množství nezbytných prováděných oprav nad soubory při konverzích, se v případě realizace převodu na straně archivátora jeví jako žádoucí vždy uchovávat vedle archivní kopie v PDF/A také originální soubory.

ZÁVĚR

Úspěšný převod více než 95,4 % nezaheslovaných souborů PDF do některé z verzí formátu PDF/A dává důvod k nepatrně většímu optimismu, než tomu bylo v případě výsledků úvodní analýzy z roku 2022, kdy úspěšnost dosahovala pouze 85,4 %. Ověření úspěšnosti konverzí do více vybraných variant formátu PDF/A potvrdilo zejména potřebu volit – na základě vlastností zdrojových PDF souborů – generačně odpovídající variantu PDF/A-1 nebo PDF/A-2, tedy něco, co použitý nástroj Adobe Acrobat automaticky neumožňuje. Ze zbylých chyb se nadále jako nejpalčivější jeví zejména problematika fontů, jejichž absence při převodu přináší významné riziko v lepším případě ztráty autenticity, v horším i přímo ztráty obsahu. Přestože tedy nadále považujeme za optimální generování archivních souborů již na straně jejich původce, budoucí konverze mezi PDF a PDF/A jsou ve většině případů realizovatelné. Z důvodu zachování autenticity nicméně doporučujeme vždy zachování také originálních souborů.

Provedená analýza zatím neposkytla jednoznačnou odpověď na otázku, zda uživatelům zpřístupňovat originální soubor PDF, nebo jeho derivát v PDF/A. Zejména s ohledem na zajištění maximální autenticity a zároveň přístupnosti (tagování) zpřístupňujících kopií dokumentů se může ukázat jako výhodnější zpřístupnění originálního souboru. Tato problematika si nicméně vyžádá podrobnější prozkoumání.

Článek vznikl v rámci institucionálního výzkumu Národní knihovny České republiky, financovaného z podpory Ministerstva kultury ČR na dlouhodobý koncepční rozvoj výzkumné organizace.

POZNÁMKY

[1] https://helpx.adobe.com/cz/acrobat/using/correcting-problem-areas-preflight-tool.html

[2] https://www.adobe.com/products/type/opentype.html; https://learn.microsoft.com/cs-cz/typography/opentype/spec/

[3] viz. Tabulka C1 https://opensource.adobe.com/dc-acrobat-sdk-docs/pdfstandards/pdfreference1.4.pdf

[4] http://martin.hoppenheit.info/blog/2018/pdfa-validation-and-inconsistent-glyph-width-information/

[5] https://github.com/veraPDF/veraPDF-validation-profiles/wiki/PDFA-Parts-2-and-3-rules#rule-66231-2

[6] https://digi.nacr.cz/validatorPDF/

LITERATURA

ANNA IRENE, Oates, 2018. *Navigating the PDF/A standard: a case study of theses in the University of Oxford's institutional repository.* Urbana, Illinois. Dostupné také z: http://hdl.handle.net/2142/100913. Master's thesis. University of Illinois at Urbana-Champaign.

FANNING, Betsy A., 2017. Preservation with PDF/A (2nd Edition). In: *DPC Technology Watch Report 17 – 01 July 2017.* 2. Great Britain: Digital Preservation Coalition, s. 1-29. ISSN 2048 -7916. Dostupné z: http://dx.doi.org/10.7207/twr17-01

FISCHER, Thomas; LUNDELL, Björn a GAMALIELSSON, Jonas. Achieving Conformance to Document Standards. Online. *International Journal of Standardization Research.* 2021, roč. 19, č. 1, s. 1-32. ISSN 2470-8542. Dostupné z: https://doi.org/10.4018/IJSR.288523. [cit. 2023-10-08].

HOPPENHEIT, Martin. PDF/A validation and inconsistent glyph width information. In: *HOPPENHEIT, Martin. Http://martin.hoppenheit.info/* [online]. 2023 [cit. 2023-10-08]. Dostupné z: http://martin.hoppenheit.info/blog/2018/pdfa-validation-and-inconsistent-glyph-width-information/

JIROUŠEK, Václav, 2018. *Formáty pro dlouhodobou archivaci elektronických publikací v rámci systému e-deposit v České republice.* Praha. Dostupné také z: http://hdl.handle.net/20.500.11956/103285. Bakalářská práce. Univerzita Karlova, Filozofická fakulta, Ústav informačních studií a knihovnictví. Vedoucí práce Souček, Martin.

JIROUŠEK, Václav a Natalie OSTRÁKOVÁ, 2017. Výběr vhodných formátů pro dlouhodobé uchování elektronických publikací. *Bulletin SKIP* [online]. 26(3) [cit. 2023-10-30]. ISSN 1213-5828. Dostupné z: https://bulletinskip.skipcr.cz/node/222

JOHNSON, Duff, 2010. Is PDF an Open Standard?. *Talking PDF* [online]. Lansdowne, Pennsylvania, U.S.A.: Appligent, ©2018, May 14, 2010 [cit. 2023-10-30]. Dostupné z: https://talkingpdf.org/is-pdf-an-open-standard/

KIRCHHOFF, Amy a Sheila MORRISSEY, 2014. Preserving eBooks. In: *DPC Technology Watch Report* [online]. Great Britain: Digital Preservation Coalition, s. 1-31 [cit. 2023-10-30]. ISSN 2048-7916. Dostupné z: doi:http://dx.doi.org/10.7207/twr14-01

KLINDT, Marco, 2017. PDF/A considered harmful for digital preservation. In: *Proceedings of the 14th International Conference on Digital Preservation, iPRES 2017, Kyoto, Japan, September 25 – 29, 2017* [online]. [cit. 2023-10-08]. Dostupné z: https://ipres2017.jp/wp-content/uploads/15Marco-Klindt.pdf

LIBRARY OF CONGRESS, 2020. *PDF/A Family, PDF for Long-term Preservation* [online]. Washington, D.C.: Library of Congress, [2020] [cit. 2023-10-08]. Dostupné z: https://www.loc.gov/preservation/digital/formats/fdd/fdd000318.shtml

LINDLAR, Michelle, Yvonne TUNNAT a Carl WILSON, 2017. A PDF Test-Set for Well-Formedness Validation in JHOVE - The Good, the Bad and the Ugly. In: *Proceedings of the 14th International Conference on Digital Preservation, iPRES 2017, Kyoto, Japan, September 25 – 29, 2017* [online]. [cit. 2023-10-30]. Dostupné z: https://ipres2017.jp/wp-content/uploads/35.pdf

MAY, Peter, 2017. Testing JHOVE PDF Module: the good, the bad, and the not well-formed. *Open Preservation Foundation* [online]. Open Preservation Foundation, 10th Mar 2017 [cit. 2023-10-30]. Dostupné z: http://openpreservation.org/blog/2017/03/10/testing-jhove-pdf-module-the-good-the-bad-and-the-not-well-formed/

MAY, Peter a Akiko KIMURA, 2019. PDF Format Preservation Assessment: Part 2: PDF/A Profile. In: *Digital Preservation Coalition Wiki* [online]. Version 1.0. British Library Digital Preservatin Team, 13 s. [cit. 2023-10-08]. Dostupné z: https://wiki.dpconline.org/images/2/22/PDFA_Assessment_v1.0.pdf

OETTLER, Alexandra. *PDF/A in a Nutshell 2.0* [online]. 1. Berlin: Association for Digital Document Standards e. V., 2013 [cit. 2023-10-08]. Dostupné z: https://pdfa.org/wp-content/uploads/2013/05/PDFA_in_a_Nutshell_211.pdf

TUNNAT, Yvonne, 2017. JHOVE – the one and only PDF validator. *Open Preservation Foundation* [online]. Open Preservation Foundation, 19th Dec 2017 [cit. 2023-10-30]. Dostupné z: http://openpreservation.org/blog/2017/12/19/jhove-the-one-and-only-pdf-validator/

Using PDF/A as a Preservation Format. In: *New York State Archives* [online]. 2013 [cit. 2023-10-08]. Dostupné z: https://www.archives.nysed.gov/records/using-pdfa-preservation-format

WHEATLEY, Paul, Peter MAY, Maureen PENNOCK, Simon WHIBLEY, David RUSSO a Akiko KIMURA, 2019. PDF Format Preservation Assessment: Part 1: PDF. In: *Digital Preservation Coalition Wiki* [online]. Version 1.5. British Library Digital Preservatin Team, 13 s. [cit. 2023-10-30]. Dostupné z: https://wiki.dpconline.org/images/f/ff/PDF_Assessment_v1.5.pdf

WOLF, Julia, 2010. *OMG WTF PDF: [PDF Ambiguity and Obfuscation]* [online]. [cit. 2023-10-30]. Dostupné z: https://www.troopers.de/media/filer_public/13/51/135142fe-0768-4b53-92a3-fd2fc7248ed2/tr11_wolf_omg_pdf.pdf

■ Článok bol recenzovaný