VÝBĚR ARCHIVAČNÍCH FORMÁTŮ PRO POVINNÝ DEPOZIT E-PUBLIKACÍ V ČR: EPUB A PDF/A JAKO ŘEŠENÍ?

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

Výběr vhodných archivačních formátů je jedním ze základních předpokladů pro dlouhodobou ochranu digitálních dokumentů. Národní knihovna ČR se dlouhodobě připravuje pro příjem povinného (či dobrovolného) výtisku textových dokumentů, vydaných v elektronické podobě. Součástí přípravných prací bylo zmapování českého trhu z pohledu struktury využívaných souborových formátů, zhodnocení těchto formátů z pohledu dlouhodobé archivace a určení vhodných nástrojů pro jejich přesnou identifikaci, charakterizaci (extrakci technických metadat) a validaci. U formátů, které jsou na trhu větší měrou zastoupeny, avšak nejsou vhodné pro dlouhodobou archivaci (zejména obecné verze formátu PDF), byly na vzorku dat ověřeny možnosti konverze do archivačních formátů PDF/A.

http://doi.org/10.52036/1335793X.2022.1-2.20-34

ÚVOD[1]

Povinný výtisk tištěných publikací plní nezastupitelnou roli ve snaze zachytit a uchovat písemné kulturní dědictví národů pro budoucí generace. Přestože se dosud v České republice nepodařilo schválit potřebné novely zákonů, které by zajistily povinné odevzdávání také elektronických publikací, již více než 10 let se Národní knihovna ČR na příjem elektronických dokumentů připravuje a sbírá zkušenosti alespoň ve formě přijímání dobrovolně odevzdaných souborů. Mezi klíčové úkoly pro zabezpečení dlouhodobé ochrany elektronických publikací patří mimo jiné dobře nastavená formátová politika pro jejich příjem i jejich následnou dlouhodobou archivaci. Nezbytným předpokladem je dobrá znalost národního trhu s e-publikacemi a jeho případných specifik. Předkládaný článek si klade za cíl přispět k poznání českého trhu nejprve prostřednictvím analýzy přidělených identifikátorů ISBN a určením aktuální struktury distribučních formátů e-publikací v České republice. U formátů, které byly identifikovány jako stěžejní pro uchování podstatné části národní produkce, je provedeno stručné zhodnocení z hlediska vhodnosti pro dlouhodobou archivaci, některých rizik pro archivaci a také dostupnosti nástrojů pro jejich zobrazení, validace nebo například konverze do jiných formátů. V následující části se článek věnuje vlastnímu vzorku zhruba 1600 e-publikací, odevzdaných mezi roky 2011 – 2021 dobrovolně do Národní knihovny ČR. Je provedena analýza složení tohoto datového souboru z hlediska formátů a kvality odevzdaných souborů. V závěrečné části se článek věnuje na vzorku 466 e-publikací praktickému ověření možností konverzí souborů PDF do archivní verze formátu PDF/A, která je jednou ze uvažovaných strategií pro dlouhodobé uložení této skupiny dat.

1. FORMÁTY E-KNIH VYUŽÍVANÉ V ČR

Elektronické publikace se mohou vyskytovat v celé řadě formátů, jejichž parametry významně předurčují možnost jejich dlouhodobého uchování. Pro vhodně nastavenou formátovou politiku sběru e-publikací je potřeba znát specifika daného národního trhu. Ke zjištění, které souborové formáty se v konkrétním čase na českém trhu s e-knihami vyskytují, se ukázala jako výhodná analýza přidělených čísel ISBN. Identifikátor ISBN je elektronickým publikacím přidělován v ČR od roku 2010 (Jeřábek, 2011). Dle pravidel pro přidělování ISBN je potom požadováno přidělení samostatného čísla ISBN vždy každému jednotlivému souborovému formátu. Při nakládání se zjištěnými daty je přesto třeba nakládat obezřetně, v praxi se lze setkat s nakladateli, kteří přidělí e-publikaci pouze jedno ISBN (např. PDF) a následně ji pod tímto identifikátorem distribuují v několika různých formátech (zpravidla PDF, EPUB, MOBI) nebo přidělí identifikátor produkčnímu formátu (např. EPUB), a následně

publikaci distribuuje pouze ve formátu jiném (např. IBOOKS). Navzdory těmto excesům ze strany nakladatelů lze považovat data za relativně spolehlivá a jiným způsobem v tomto rozsahu nezjistitelná.
Analýza databáze ohlášených e-knih české agentury ISBN byla provedena v listopadu 2021, kdy obsahovala na 100 443 evidovaných čísel ISBN od více než 2 000 nakladatelů. Jako nejpočetnější za celé sledované období 2010 – 2021 se ukázal formát EPUB se 40,6 %, formát PDF s 36,1 %, formát MOBI s 21,1 % a HTML s 1,3 %. Na ostatní formáty zbylo 0,9 % přidělených identifikátorů ISBN.

Pro zjištění aktuální podoby trhu s e-knihami jsme provedli podrobnější analýzu posledního ukončeného roku 2020. Pro lepší představu o významu jednotlivých formátů v kontextu celkové produkce byly mezi sledované ukazatele vedle míry zastoupení doplněny také počty nakladatelů, kteří v daném formátu v roce 2020 publikovali a počet titulů, které byly vydány pouze v tomto formátu.
Celkem bylo v roce 2020 přiděleno e-publikacím 9 050 čísel ISBN patřícím k 577 nakladatelům. Uvedený počet ISBN náležel podle porovnání jedinečných názvů přibližně 5 096 různým titulům, z nichž 2736 vyšlo

formát	EPUB	PDF	MOBI/PRC	HTML	WOOKY	IBOOKS	ANDROID	iOS	ostatní
počet	40 817	36 228	21 243	1 258	222	132	67	67	409
%	40,6 %	36,1 %	21,1 %	1,3 %	0,2 %	0,1 %	0,1 %	0,1 %	0,4 %

Tabulka 1. Přehled formátů e-publikací ohlášených agentuře ISBN za období 2010 – listopad 2021

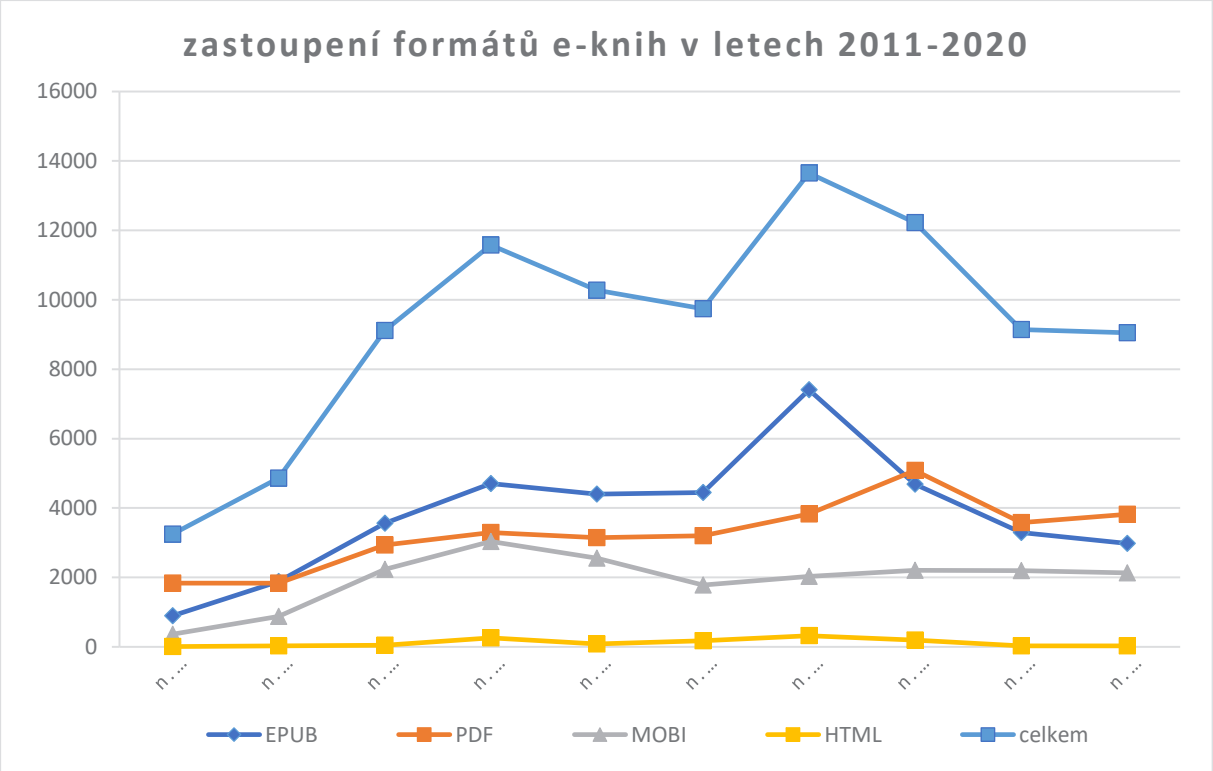
formát	PDF	EPUB	MOBI/PRC	HTML	HTML5	AZW3	ostatní
počet	3858	2980	2131	29	23	19	10
%	42,6 %	32,9 %	23,5 %	0,3 %	0,3 %	0,2 %	0,1 %
počet titulů vydaných pouze v daném formátu	1982	632	65	29	23	1	5
nakladatelů	506	252	174	5	1	5	9

Tabulka 2. Přehled formátů e-publikací ohlášených agentuře ISBN za rok 2020

pouze v jednom formátu, 779 ve dvou formátech (nejčastěji v kombinaci EPUB+MOBI, nebo EPUB+PDF), 1566 ve třech formátech (téměř vždy v kombinaci PDF+EPUB+MOBI/PRC) a 10 titulů ve čtyřech či více formátech. V absolutních číslech byl nejvíce zastoupeným formátem PDF s 42,6 %, druhým v pořadí EPUB s 32,9 % a MOBI s 23,5 % (MOBI+PRC). Formáty PDF, EPUB a MOBI tedy v roce 2020 tvořily v absolutních číslech rovných 99 % všech vydaných e-knih v České republice. Přestože pak byl formát MOBI zastoupen celkem 23,5 %, jeho význam z pohledu jeho potřeby pro dlouhodobé uchování výrazně poklesne při pohledu na počet titulů, publikovaných pouze jako MOBI. 96,5 % titulů vydaných v tomto formátu má svůj ekvivalent v jiném formátu, nejčastěji v PDF nebo EPUB, které, jak posléze ukážeme, se nám jeví z pohledu dlouhodobé archivace jako perspektivnější.
Z pohledu analýzy zastoupení formátu spíše okrajovým jevem je patrný, téměř 30% pokles počtu celkově přidělených ISBN v letech 2019 a 2020 oproti předchozímu období 2017 a 2018. Aktuální data za rok 2021 nebyla v době analýzy k dispozici, setrvalost trendu tedy nebylo zatím možné potvrdit.

2. VÝBĚR FORMÁTŮ PRO DLOUHODOBOU ARCHIVACI

Výběr vhodných archivačních formátů představuje zásadní krok pro dlouhodobé uchování digitálních dokumentů. Při výběru archivačních formátů je třeba tyto formáty posuzovat podle celé řady kritérií, která předurčují jejich vhodnost pro dlouhodobou archivaci. Mezi nejdůležitějšími se zpravidla uvádí rozšířenost, otevřenost a nezatíženost patenty – tedy dostupnost formátové specifikace, dostatečná velikost uživatelské komunity a z ní vyplývající dostupnost nástrojů pro práci s daným formátem a také to, zda užití formátu



Obrázek 1. Počet přidělených ISBN jednotlivým formátům v letech 2011 – 2020

není svázáno licenčními omezeními. K dalším kritériím můžeme zařadit například standardizaci formátu některou z uznávaných normalizačních organizací (např. ISO), míru nezávislosti na specifických hardwarových a softwarových platformách, nebo například možnost interoperability s jinými formáty. (Cubr, 2010, s. 83 – 86)

Pokud má zvolený archivní formát sloužit jako cílový formát konverzí z původních produkčních či distribučních formátů, měl by zároveň být schopen plnit roli tzv. archivní matrice, tedy umožnit zachovat nejvyšší možnou kvalitu zdrojových dokumentů a uchovat všechny signifikantní vlastnosti[2], jejichž uchování jsme si definovali za nezbytné pro zachování autenticity archivovaného dokumentu (Cubr, 2017). Při výběru archivačních formátů je dobrou praxí řídit se doporučeními mezinárodně uznávaných institucí (Ostráková a Kopský, 2020). Mezi základní zdroje informací pro hodnocení a doporučení vhodných formátů pro dlouhodobou archivaci dle jednotlivých typů dokumentů tak patří zejména každoročně aktualizovaný formátový registr Kongresové knihovny (Library of Congress, 2021), v oblasti elektronických publikací je cenným zdrojem informací například hodnocení formátů Britské národní knihovny (British Library, 2021).

FORMÁT EPUB

Formát EPUB byl v roce 2007 vydán organizací International Digital Publishing Forum. Od počátku se jedná o otevřený formát, založený zejména na standardech XML, XHTML, CSS a formátu ZIP. Formát byl vytvořen jako nástupce staršího Open eBook Publication Structure (1999), první verze EPUB byla proto označena rovnou číslem 2.0. V roce 2010 vyšla revize formátu 2.0.1 a v roce 2011 byl vydán nástupnický formát EPUB 3.0, postupně aktualizovaný na verze 3.0.1 (2014), 3.1 (2017) a 3.2 (2019). Formát EPUB 3 je komplexnější než EPUB 2, umožnuje například fixní sazbu či lépe definovat náročnější grafickou úpravu publikací, přidává podporu JavaScriptu nebo vkládání multimediálního obsahu. Formát EPUB 3 jakožto otevřený standard slouží také tako základ pro řadu odvozených formátů, jako jsou například iBooks od firmy Apple, KF8 od Amazonu nebo v českém prostředí formát Wooky od stejnojmenné firmy. Odvozené proprietární formáty však již zpravidla obsahují řadu proprietárních rozšíření, které zpravidla přinejmenším brání zobrazení na zařízeních jiných výrobců. (Day, May, Wheatley, Whibley a Knijff, 2019)

Specifikace pro EPUB 3.0 byla v roce 2014 Mezinárodní organizací pro normalizaci (ISO) vydána jako norma

ISO/IEC TS 30135 (části 1 – 7). EPUB 3.0.1 byl v roce 2020 vydán jako ISO/IEC 23736 (části 1 – 6). Vedle samotné existence standardizace formátu na úrovni mezinárodních technických norem ISO lze pro dlouhodobou ochranu považovat rovněž v roce 2020 vydanou normu ISO/IEC TS 22424 (části 1–2), která obsahuje konkrétní doporučení pro tvorbu souborů ve formátu EPUB 3, s ohledem na dlouhodobou uchovatelnost obsahu. Formát EPUB vykazuje díky velkému rozšíření poměrně dobrou podporu aplikací pro zobrazení, validace i konverze vůči jiným formátům. K identifikaci formátu (nikoliv však verze) lze využít služeb nástroje DROID, k určení konkrétní verze formátu, validaci a extrakci technických metadat lze využít nástroj EPUBCheck, pro převody je k dispozici například nástroj Calibre.

Formát EPUB lze považovat za relativně vhodný pro dlouhodobou archivaci, má však několik vlastností, které je zejména u verze EPUB 3 nezbytné sledovat a přijímat opatření, která mohou v dlouhodobé perspektivě předcházet ztrátě obsahu. K problematickým funkcím formátu patří, například možná závislost na externím (např. multimediálním) obsahu, podpora JavaScriptu nebo přítomnost DRM ochrany, včetně podpory šifrování fontů (font obfuscation). (Knijff, 2012) Bezpečné dlouhodobé uchovávání souborů EPUB a autenticitu jejich obsahu proto nelze garantovat, pokud nejsou rizika u jednotlivých souborů včas identifikována a přijata opatření k jejich nápravě. Při příjmu souborů EPUB k dlouhodobému uložení lze proto doporučit nejen formální validaci vůči formátové specifikaci, ale také detekovat konkrétní vlastnosti, které jsou sice v souladu se standardem, ale v rozporu s principy dlouhodobé archivace (Knijff, 2015). Tyto vlastnosti je třeba definovat jako součást formátové politiky instituce a buď odevzdávání nevhodně vytvořených souborů od nakladatelů předcházet, nebo provádět nápravná opatření po odevzdání těchto dat a v těch případech, kdy nápravná opatření nejsou možná, jednoznačně deklarovat, u kterých vlastností původních souborů (např. přehrání externě odkazovaného multimediálního obsahu) není dlouhodobá ochrana garantována.

FORMÁT MOBI

Formát MOBI byl vytvořen francouzskou firmou Mobipocket SA krátce po jejím vzniku v březnu 2000. Cílem bylo vytvořit formát, který umožní formátování textu a bude vhodný pro tehdy rozšířená zařízení PDA. Po koupi firmy americkou společností Amazon v roce 2005 byl formát spojen zejména s její čtečkou Kindle. Díky oblíbenosti hardwarových čteček Kindle, které tradičně uměly zobrazovat právě soubory MOBI, nikoliv EPUB, se hojně rozšířil také v České republice. Přestože tedy internetový obchod Amazonu e-knihy v českém jazyce ve své nabídce nenabízí, stal se MOBI na českém trhu oblíbeným distribučním formátem. V roce 2007 vytvořil Amazon nástupnický, lehce modifikovaný formát AZW, který začal být po roce 2011 nahrazován formátem KF8, odvozeným od EPUB 3. V roce 2011 oznámil Amazon ukončení podpory formátu MOBI a v roce 2016 došlo navíc k ukončení provozu webu Mobipocket.com, obsahujícího (nekompletní) část publikované dokumentace formátu MOBI. V dubnu 2022 oznámil Amazon ukončení podpory formátu MOBI v aplikaci Kindle Library v průběhu roku 2022 a zahájení podpory formátu EPUB (Schroeder, 2022). V budoucnu lze proto očekávat další pokles významu formátu nejen na českém trhu ve prospěch otevřeného standardu EPUB. (Jiroušek, 2018)

Podobně jako EPUB byl formát MOBI vytvořen na základě formátu Open eBook Publication Structure (1999) a dále formátu PalmDOC (1996). Formát je založen především na standardech XHTML a CSS, binární formát sloužící k zabalení a distribuci byl původně odvozen od formátu PalmDOC. Od formátu PalmDOC také zdědil variantní příponu .prc, která se někdy používá namísto běžnější přípony .mobi. Částečná specifikace formátu MOBI, která se týká definice obsahu, jeho formátování a připojených metadat byla do roku 2016 dostupná na webu Mobipocket.com (dnes je dostupná prostřednictvím služby Internet Archive's Wayback machine). Specifikace binárního formátu oproti tomu nebyla nikdy oficiálně publikována a nekompletní dokumentaci se tak podařilo získat jedině metodou reverzního inženýrství. (Day, Russo a Whibley, 2019) Nedostupnost kompletní dokumentace formátu a s tím související absence validačních nástrojů, společně s morálním zastaráváním a klesající podporou čtecích zařízení činí z MOBI formát, který představuje z pohledu dlouhodobé archivace značné riziko. K dispozici jsou sice nástroje, které umožňují konverzi do otevřených formátů (např. nástroj Calibre), vzhledem k nekompletní dokumentaci formátu však nemusí ani tyto nástroje pracovat spolehlivě.

FORMÁT PDF

Formát PDF byl společností Adobe Systems představen v roce 1993. Jeho specifikace ve velké míře vycházela z programovacího jazyka Post Script a hlavním úkolem nového formátu bylo zajistit jednotné zobrazení dokumentů nezávisle na použitém software, hardware a operačním systému. Do roku 2006 postupně firma Adobe vydala celkem sedm revizí standardu: 1.1 (1994), 1.2 (1996), 1.3 (2000), 1.4 (2001), 1.5 (2003),

1.6 (2004) a 1.7 (2006). Další vývoj standardu pak přešel pod Mezinárodní organizaci pro normalizaci, která nejprve vydala v roce 2008 verzi PDF 1.7 jako normu ISO 32000-1:2008 a v roce 2017 také zatím poslední velkou revizi formátu PDF 2.0 jako ISO 32000-2:2017.
Do rodiny formátů dnes patří PDF a řada dalších odvozených standardů s vlastními názvy, jen namátkou uveďme PDF/H pro zdravotnictví, PDF/X pro tisk, PDF/UA pro zrakově postižené nebo PDF/A pro archivaci. Celkový počet verzí jednotlivých formátů PDF dosahuje mnoha desítek. Za nejdůležitější z pohledu dlouhodobé archivace můžeme nepochybně považovat standardy PDF/A, existující dnes ve čtyřech základních verzích PDF/A-1 (2005), PDF/A-2 (2011), PDF/A-3 (2012) a PDF/A-4 (2020). Všechny verze PDF/A byly vydány jako ISO standardy. Jednotlivé verze PDF/A jsou fakticky formátovými profily PDF, které zakazují využití jinak podporovaných technologií a formátů, které byly technickou komisí ISO považovány za rizikové z hlediska dlouhodobé archivace (např. audiovizuální objekty, 3D objekty, JavaScript nebo šifrování) a striktně vyžadují uložení veškerých informací a dat potřebných k vykreslení obsahu (jako jsou například fonty) přímo do souboru PDF/A. PDF/A-1 je profilem z PDF 1.4, PDF/A-2 a PDF/A-3 jsou odvozeny od PDF 1.7 a PDF/A-4 z PDF 2.0. Od verze PDF/A-3 je podporováno také vkládání libovolných dalších souborových formátů do souboru PDF/A, garantováno je pak nicméně zachování pouze vizuální reprezentace takto vloženého obsahu. (Fanning, 2017) (Kimura a May, 2019)
Při posuzování formátu PDF z pohledu dlouhodobé archivace narážíme na několik problémů. Navzdory dostupnosti specifikace přináší komplikace jednak značná komplexnost formátu a také určitá vágnost formátové specifikace u nejstarších verzí formátu PDF. Přílišná komplexnost formátu znesnadňuje jednak plnou implementaci standardů v nástrojích pro tvorbu, zpřístupnění nebo validaci. Někdejší nejednoznačnosti formátové dokumentace se navíc Adobe rozhodlo čelit pragmaticky tím, že jeho vlastní referenční implementace PDF, proprietární aplikace Adobe Reader, dokáže zobrazit i velmi nestandardní a nevalidní soubory PDF. Vzhledem k absenci efektivních validačních nástrojů se pak historicky stala zobrazitelnost souborů v proprietárním Adobe Readeru hlavním cílem vývojářů nástrojů pro tvorbu PDF (Johnson, 2010). Soubor, který tedy dnes například bez problému otevřeme v aplikaci Adobe Acrobat Readeru DC, se v jiné aplikaci pro zobrazení PDF nemusí zobrazit správně. Prakticky jediným validačním nástrojem pro validaci PDF souborů je aplikace JHOVE a její modul PDFhul. Vzhledem k výše popsaným problémům je však třeba počítat s nedokonalostí výstupů validace. Lepší situace je u PDF/A, kde je k dispozici validační nástroj veraPDF. Oba open-source nástroje jsou spravovány a rozvíjeny organizací Open Preservation Foundation a jejich výhodou je proto silná a aktivní uživatelská komunita. (Wheatley, May, Pennock, Whibley, Russo a Kimur, 2019)
Hlavní rizika formátu PDF pro dlouhodobou archivaci lze tedy shrnout jako přílišnou komplexnost formátu, možnost vkládání prakticky libovolného dalšího souborového formátu (jehož zobrazitelnost v budoucnu nelze jednoduše garantovat) – zejména možnost vkládání audiovizuálního obsahu, 3D objektů nebo JavaScriptu. Riziko může dále představovat přítomnost DRM, závislost na externím obsahu nebo také přítomnost malware. Všechna zmíněná rizika jsou řešena v rámci profilů PDF/A, které je výrazně minimalizují. Určitou nevýhodou formátu PDF/A je naopak omezená schopnost sloužit jako archivní matrice, neboť určitý typ obsahu do něj není možné z principu bez ztráty převést. Mnohdy pak nelze PDF/A vytvořit z legálních důvodů, kdy například licenční politika znemožní vložit do souboru potřebné fonty, případně nejsou tyto fonty při převodu k dispozici. Nahrazování fontů pak může vést nejen ke ztrátě autenticity, ale také ke ztrátě informačního obsahu.

3. ANALÝZA VZORKU E-KNIH ODEVZDANÝCH DO NK ČR

Dobrovolný příjem elektronických publikací zahájila Národní knihovna ČR ve zkušebním režimu poprvé v roce 2011 v rámci projektu eDpilot (2011 – 2012) a víceméně nepřetržitě probíhá dosud. Odevzdávání e-publikací probíhá převážně prostřednictvím e-mailu, publikace zpřístupněné v režimu open access jsou stahovány přímo z webových stránek vydavatelů, méně časté je předávání na fyzických nosičích nebo prostřednictvím různých webových služeb pro výměnu dat. Národní knihovna ČR se v minulosti zavazovala k dlouhodobé ochraně těch publikací, které budou odevzdány ve formátech PDF/A-1, PDFA-2 nebo EPUB 2. Nakladatelé tak mohli být z části motivováni odevzdávat data především v těchto formátech. Ostatní souborové formáty byly přijímány pouze k bitové ochraně a přestože je zájmem Národní knihovny i tyto dokumenty dlouhodobě ochránit, jejich dlouhodobou dostupnost nelze garantovat. V současné době disponuje Národní knihovna ČR vzorkem přibližně 1 600 takto odevzdaných dokumentů. Z pohledu celkové produkce e-publikací v ČR se nejedná o zcela reprezentativní vzorek, představuje však poměrně pestrou mozaiku souborů od několika desítek převážně menších nakladatelů a jako takový může vhodně dokreslit

představu o českém trhu s epublikacemi ve větším detailu, nežli umožňuje pohled na databázi přidělených ISBN.

Analyzovaný vzorek dat čítal přesně 1585 datových souborů. Pro úvodní identifikaci souborových formátů byl použit nástroj Droid (v6.5.2), kterým bylo ve vzorku zjištěno zejména 1024 souborů EPUB, 473 souborů PDF a 73 souborů MOBI. Pro určení konkrétních verzí formátu EPUB bylo nutné využít modul EPUB-ptc (Release 1.0) v rámci nástroje JHOVE, identické výsledky poskytl rovněž EPUBCheck.

Nejsilněji zastoupeným souborovým formátem ve vzorku byl s 1024 soubory formát EPUB. V případě tohoto souborového formátu je nicméně třeba uvést, že plných 1000 souborů pochází z produkce jediného nakladatele a poněkud tak narušuje rozmanitost a reprezentativnost vzorku. K validaci souborů byl využit nástroj EPUBCheck (v. 4.2.6). Validace určila 1006 souborů jako validních, výsledek u 4 souborů obsahoval varování a u 13 souborů ohlásil EPUBCheck chybu. U všech nevalidních souborů se potvrdila zobrazitelnost aplikací Calibre bez viditelných chyb. U žádného z kontrolovaných souborů také nebyla zjištěna přítomnost DRM. Oprava zjištěných chyb nebyla předmětem analýzy.
Soubory ve formátu MOBI byly zastoupeny 73 exempláři. Z tohoto počtu má 20 souborů koncovku .mobi a 53 souborů koncovku .prc. Vzhledem k absenci validačních nástrojů bylo u této skupiny dokumentů možné ověřit pouze zobrazitelnost v dostupných nástrojích. U všech 53 souborů s koncovkou .prc byla zjištěna přítomnost DRM zabezpečení, která brání zobrazitelnosti mimo oficiální (a nyní již oficiálně nedostupnou) aplikaci Mobipocket a rovněž převodu do jiných formátů. Zbylých 20 souborů s koncovkou .mobi je například v aplikaci Calibre bez omezení zobrazitelných.

koncovka	PUID	název formátu	verze formátu	počet souborů
epub	fmt/483	ePub format	2.0.1	1018
pdf	fmt/19	Acrobat PDF 1.5 – Portable Document Format	1.5	142
pdf	fmt/20	Acrobat PDF 1.6 – Portable Document Format	1.6	116
pdf	fmt/18	Acrobat PDF 1.4 – Portable Document Format	1.4	116
pdf	fmt/276	Acrobat PDF 1.7 – Portable Document Format	1.7	59
prc	fmt/396	PocketMobi (Palm Resource) File		53
pdf	fmt/17	Acrobat PDF 1.3 – Portable Document Format	1.3	36
mobi	fmt/396	PocketMobi (Palm Resource) File		20
pdf	fmt/503	AppleDouble Resource Fork	2.1	9
epub	fmt/483	ePub format	3.2	6
pdf	fmt/16	Acrobat PDF 1.2 – Portable Document Format	1.2	4
doc	fmt/40	Microsoft Word Document	97-2003	3
ibooks	fmt/482	Apple iBook format		1
docx	fmt/412	Microsoft Word for Windows	2007 onwards	1

Tabulka 3. Výsledek identifikace souborových formátů aplikací Droid (verze EPUB určeny pomocí JHOVE)

Soubory EPUB: výsledek kontroly nástrojem EPUBCheck	počet souborů	%
Error	14	1,4 %
Valid	1006	98,2 %
Warrning	4	0,4 %

Tabulka 4. Výsledek kontroly souborů EPUB aplikací EPUBCheck

Soubory EPUB: ID chyby z nástroje EPUBCheck	klasifikace	počet výskytů
NCX-001	ERROR	7
RSC-005	ERROR	149
PKG-006	ERROR	2
RSC-001	ERROR	2
RSC-003	ERROR	1
RSC-012	ERROR	10
RSC-002	FATAL	2
PKG-008	FATAL	2
PKG-004	FATAL	1
RSC-004	INFO	10
CSS-007	INFO	5
OPF-055	WARNING	1
PKG-010	WARNING	18
HTM-014	WARNING	38
OPF-003	WARNING	2

Tabulka 5. Seznam zjištěných chyb v souborech EPUB, ohlášených aplikací EPUBCheck

Soubory PDF: JHOVE status	počet souborů	%
Not Well-Formed	57	12,2 %
Well-Formed and valid	399	85,6 %
Well-Formed but not valid	10	2,1 %

Tabulka 6. Výsledek kontroly zdrojových PDF souborů v aplikaci JHOVE (PDF-hul)

Soubory PDF: výstupy validace nástrojem JHOVE (ErrorMesage)	ID	počet výskytů
Compression method is invalid or unknown to JHOVE	PDF-HUL-109	23
Invalid character in hex string	PDF-HUL-11	1
Invalid object number or object stream	PDF-HUL-110	5
Malformed outline dictionary	PDF-HUL-124	4
Malformed dictionary: Vector must contain an even number of objects, but has 19	PDF-HUL-41	6
Malformed dictionary	PDF-HUL-42	6
Malformed filter	PDF-HUL-45	1
Lexical error	PDF-HUL-66	2
Invalid Names dictionary	PDF-HUL-90	3
Unexpected exception java.lang.NullPointerException	PDF-HUL-94	7

Tabulka 7. Seznam zjištěných chyb ve zdrojových PDF souborech, ohlášených aplikací JHOVE (PDF-hul)

Zkušební převod části těchto 20 souborů do formátu EPUB vedl k vytvoření sice zobrazitelných, nicméně nevalidních souborů. Automatická konverze souborů MOBI do EPUB je tedy teoreticky možná, nelze nicméně očekávat takovou spolehlivost, aby bylo možné je využívat bez výraznějších ručních zásahů a kontrol.
Z hlediska velikosti a rozmanitosti lze za nejzajímavější část vzorku e-publikací považovat 473 souborů PDF. Po odstranění duplicitních a jiných souborů z adresářové struktury (dopis autora apod.) bylo k testování vybráno 466 jednotlivých souborů PDF. Identifikace souborových formátů nástrojem Droid ve vzorku určila šest verzí formátu PDF, od verze 1.2 do verze 1.7. Nástroj JHOVE oproti Droidu zprvu nerozpoznal 60 souborů PDF, po vynucené validaci modulem PDF-hul specializovaným na validaci PDF souborů, nedokázal stále určit verzi PDF u 41 souborů. 57 souborů (12,2 % z celku) z 60 původně nerozpoznaných určil následně JHOVE jako Not well-formed. Jako Well-formed and valid bylo naopak určeno 399 souborů (85,6 %), zbylých 10 souborů bylo označeno jako Well-formed but not valid.

4. ZKUŠEBNÍ PŘEVOD PDF DO PDF/A

K převodu PDF do PDF/A-1b byl využit Adobe Acrobat Pro DC (verze 2022.001.20117) ve výchozím nastavení. 6 souborů již bylo ve formátu PDF/A a 15 souborů bylo chráněno heslem proti úpravám a nebylo je možné převést. U 17 souborů aplikace převod nebyla schopna dokončit (z toho 9 souborů bylo původně identifikováno jako nevalidních aplikací JHOVE), naopak u 428 souborů úspěšně proběhl. Všechny úspěšně vytvořené soubory byly následně validovány validátorem VeraPDF (verze 1.20.1) pro ověření validnosti souborů vůči odpovídajícímu profilu PDF/A. Celkem 374 vytvořených souborů PDF/A bylo vyhodnoceno jako validní, naopak 54 souborů bylo označeno jako nevalidní. Úspěšnost konverze tedy dosáhla přibližně 81,3 % (374 ze 460). Úspěšně převést do podoby validních PDF/A-1b souborů se podařilo také 38 z 67 původně nevalidních souborů PDF, tedy 56,7 %. Konkrétní chybová hlášení při konverzích v nástroji Adobe Acrobat i při validaci v nástroji VeraPDF zachycují tabulky 9 a 10.
U všech převedených souborů PDF/A (včetně nevalidních) bylo následně provedeno také vizuální porovnání proti zdrojovým PDF souborům. Porovnání probíhalo otevřením souborů v Adobe Acrobat a vizuálním porovnáním několika prvních stránek dokumentů. Z 374 úspěšně převedených a validních souborů PDF/A byly u 7 souborů (1,9 %) zjištěny vizuální rozdíly oproti zdrojovým souborům (viz obrázky 2 až 7). Vizuální rozdíly se ve všech sedmi případech týkaly snížení kvality obrazů připomínající zvýšení komprese, u žádného z validních dokumentů nebyla touto metodou zjištěna například změna fontu nebo chybějící znaky. Výrazně jiná situace byla u těch souborů, které byly sice úspěšně převedeny, ale jež VeraPDF označila jako nevalidní. Z 54 těchto souborů byl rozdíl zjištěn hned v 26 případech (48,1 %). Vedle snížení kvality obrazů u těchto dokumentů v řadě případů došlo, zřejmě v důsledku nedostupnosti fontů, také ke ztrátě některých znaků či rozsáhlejších částí textu (Obr. 8 až 15).

ZÁVĚR

Úvodní analýza zastoupení formátů na českém trhu s e-publikacemi potvrdila význam formátů PDF a EPUB. Význam formátu MOBI naopak potvrzen nebyl a aktuální vývoj okolo tohoto formátu naznačuje jeho postupný ústup nejen z českého trhu ve prospěch otevřeného formátu EPUB. Rizika, která formát MOBI může z pohledu dlouhodobé archivace přinášet, potvrdilo dílčím způsobem i kvalitativní posouzení vzorku dat, tvořeným e-publikacemi dobrovolně odevzdanými do Národní knihovny ČR mezi lety 2011 – 2021. 53 ze 73 odevzdaných souborů MOBI nebylo možné v dnes běžně dostupných nástrojích zobrazit a z důvodu přítomnosti DRM ochrany ani převést do jiných formátů, zbylých 20 souborů se zobrazit podařilo, jejich převod do podoby validních souborů EPUB nicméně úspěšný nebyl.
Přítomnost ochrany, bránící konverzím do jiného formátu, byla detekována také u 15 souborů PDF, u nichž nebylo možné realizovat konverzi do PDF/A. Naopak, úspěšně převést do podoby validních PDF/A souborů se podařilo větší část z 67 publikací, které byly původně ve formátu PDF vyhodnoceny validačním nástrojem JHOVE jako Not-well formed. Provedené převody a vyhodnocení následných validací v kombinaci s vizuálními kontrolami převedených dokumentů přinesly několik poznatků. Riziko ztráty informačního obsahu při konverzi PDF do PDF/A, způsobené nejčastěji nedostupností fontů v cílovém PDF/A se potvrdilo jako skutečnost, se kterou je třeba při archivaci PDF souborů počítat. Zjištění vizuálních rozdílů u přibližně 2 % úspěšně převedených a validních souborů PDF/A ukazuje potřebu na jednoznačnou formulaci těch signifikantních vlastností, které požadujeme zachovat v průběhu jakýchkoliv formátových změn v současnosti i v budoucnosti. Vedle jednoznačné definice těchto vlastností je zapotřebí také určit vhodné nástroje pro jejich automatickou kontrolu. Jako pozitivní lze hodnotit úspěšnost validací nástrojem VeraPDF z pohledu zachycení těch souborů, u nichž došlo v důsledku kon-

Výsledky převodu PDF do PDF/A	počet	%
originální soubor je PDF-A	6	1,2 %
validní po převodu	374	80,2 %
nevalidní po převodu	54	11,6 %
nejde převést z důvodu ochrany	15	3,2 %
nepřevedeno	17	3,6 %

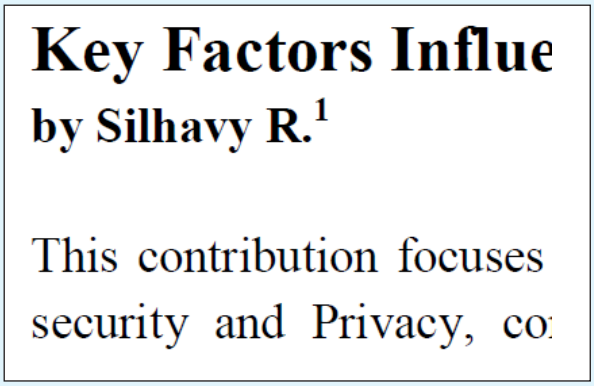
Tabulka 8. Úspěšnost převodů PDF do PDF/A

Hlášení chyb při konverzi (Adobe Acrobat)	počet výskytů
EmbeddedFiles entry in Names dictionary	7
Glyphs missing in embedded font	11
CharSet incomplete for Type 1 font	6
PDF contains EF	1
PDF contains EF (embedded file) entry	6
Syntax problem: Array with more than 8191 elements	6
Syntax problem: Dictionary with more than 4095 entries	1
Transparency used (transparency group)	1
Width information for rendered glyphs is inconsistent	11

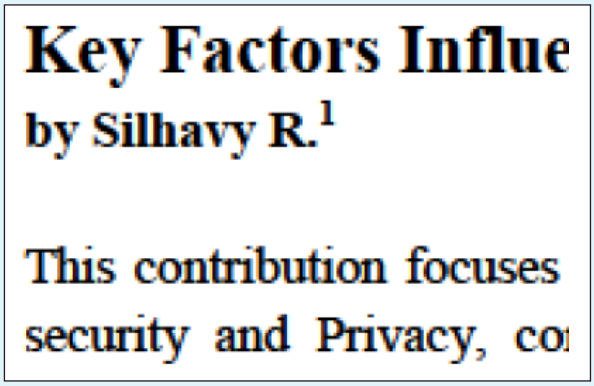
Tabulka 9. Seznam chyb ohlášených aplikací Adobe Acrobat u nedokončených konverzí PDF do PDF/A

Hlášení chyb při validaci (VeraPDF)	počet výskytů
A file specification dictionary, as defined in PDF 3.10.2, shall not contain the EF key	7
A file's name dictionary, as defined in PDF Reference 3.6.3, shall not contain the EmbeddedFiles key	24
A Group object with an S key with a value of Transparency shall not be included in a form XObject. A Group object with an S key with a value of Transparency shall not be included in a page dictionary	1
DeviceCMYK may be used only if the file has a PDF/A-1 OutputIntent that uses a CMYK colour space	1
Embedded font programs shall define all font glyphs referenced for rendering with conforming file.	5
For all Type 1 font subsets referenced within a conforming file, the font descriptor dictionary shall include a CharSet string listing the character names defined in the font subset, as described in PDF Reference Table 5.18	6
For every font embedded in a conforming file and used for rendering, the glyph width information in the font dictionary and in the embedded font program shall be consistent.	17
Maximum capacity of a dictionary (in entries) is 4095	1
Maximum capacity of an array (in elements) is 8191	6
The PDF/A version and conformance level of a file shall be specified using the PDF/A Identification extension schema.	28

Tabulka 10. Hlášení chyb při validaci (VeraPDF)



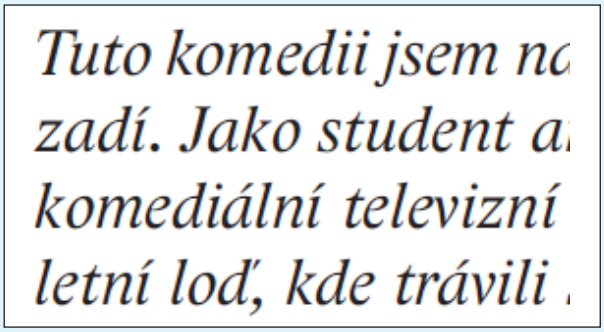
Obrázek 2. Původní soubor PDF před konverzí



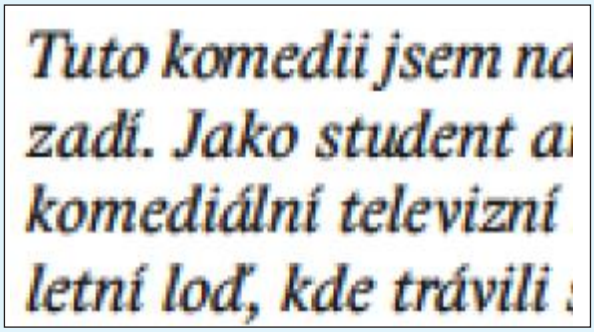
Obrázek 3. Výsledný soubor po převodu do PDF/A (validní)

Obrázek 4. Původní soubor PDF před konverzí

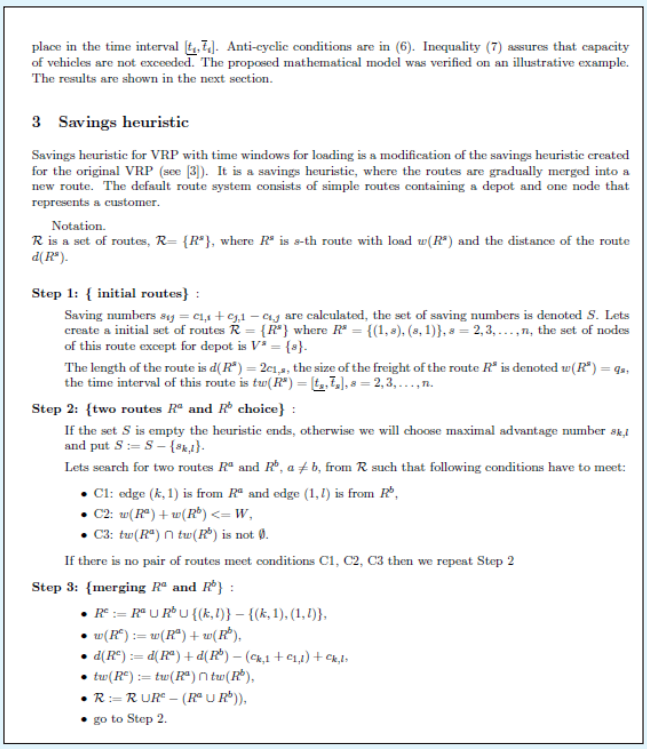
Obrázek 5. Výsledný soubor po převodu do PDF/A (validní)



Obrázek 6. Původní soubor PDF před konverzí



Obrázek 7. Výsledný soubor po převodu do PDF/A (validní)



place in the time interval $[\underline{t_i}, \overline{t_i}]$. Anti-cyclic conditions are in (6). Inequality (7) assures that capacity of vehicles are not exceeded. The proposed mathematical model was verified on an illustrative example. The results are shown in the next section.

3 Savings heuristic

Savings heuristic for VRP with time windows for loading is a modification of the savings heuristic created for the original VRP (see [3]). It is a savings heuristic, where the routes are gradually merged into a new route. The default route system consists of simple routes containing a depot and one node that represents a customer.

Notation.
$\mathcal{R}$ is a set of routes, $\mathcal{R}$= $\{R^s\}$, where R^s is s-th route with load $w(R^s)$ and the distance of the route $d(R^s)$.

Step 1: { initial routes} :

Saving numbers $s_{ij} = c_{1,i} + c_{j,1} - c_{i,j}$ are calculated, the set of saving numbers is denoted S. Lets create a initial set of routes $\mathcal{R} = \{R^s\}$ where $R^s = \{(1,s),(s,1)\}, s = 2,3,\ldots,n$, the set of nodes of this route except for depot is $V^s = \{s\}$.

The length of the route is $d(R^s) = 2c_{1,s}$, the size of the freight of the route R^s is denoted $w(R^s) = q_s$, the time interval of this route is $tw(R^s) = [\underline{t_s}, \overline{t_s}], s = 2,3,\ldots,n$.

Step 2: {two routes R^a and R^b choice} :

If the set S is empty the heuristic ends, otherwise we will choose maximal advantage number $s_{k,l}$ and put $S := S - \{s_{k,l}\}$.

Lets search for two routes R^a and R^b, $a \neq b$, from $\mathcal{R}$ such that following conditions have to meet:

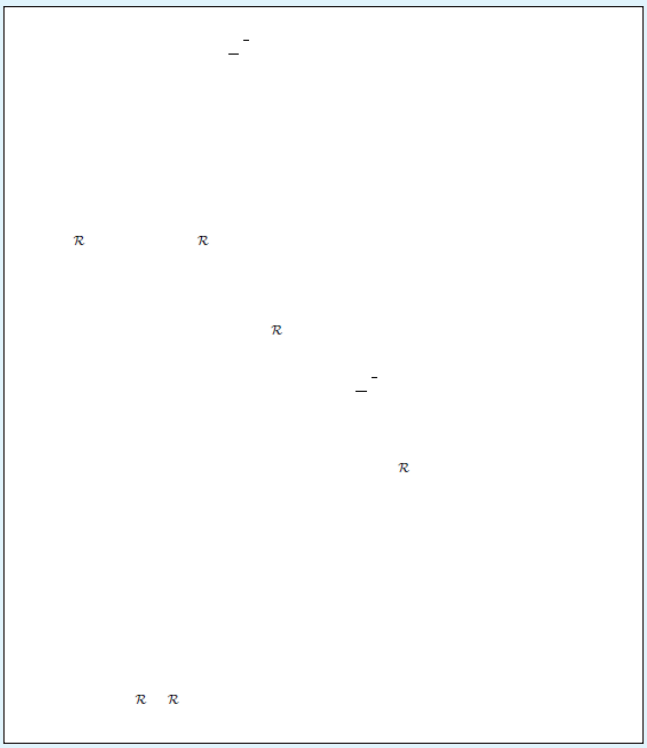
- C1: edge $(k,1)$ is from R^a and edge $(1,l)$ is from R^b,
- C2: $w(R^a) + w(R^b) <= W$,
- C3: $tw(R^a) \cap tw(R^b)$ is not $\emptyset$.

If there is no pair of routes meet conditions C1, C2, C3 then we repeat Step 2

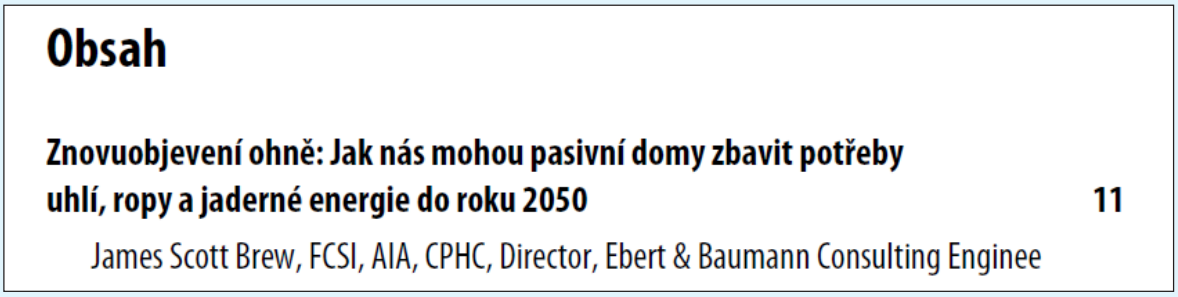
Step 3: {merging R^a and R^b} :

- $R^c := R^a \cup R^b \cup \{(k,l)\} - \{(k,1),(1,l)\}$,
- $w(R^c) := w(R^a) + w(R^b)$,
- $d(R^c) := d(R^a) + d(R^b) - (c_{k,1} + c_{1,l}) + c_{k,l}$,
- $tw(R^c) := tw(R^a) \cap tw(R^b)$,
- $\mathcal{R} := \mathcal{R} \cup R^c - (R^a \cup R^b))$,
- go to Step 2.

Obrázek 8. Původní soubor PDF před konverzí



Obrázek 9. Výsledný soubor po převodu do PDF/A (nevalidní)

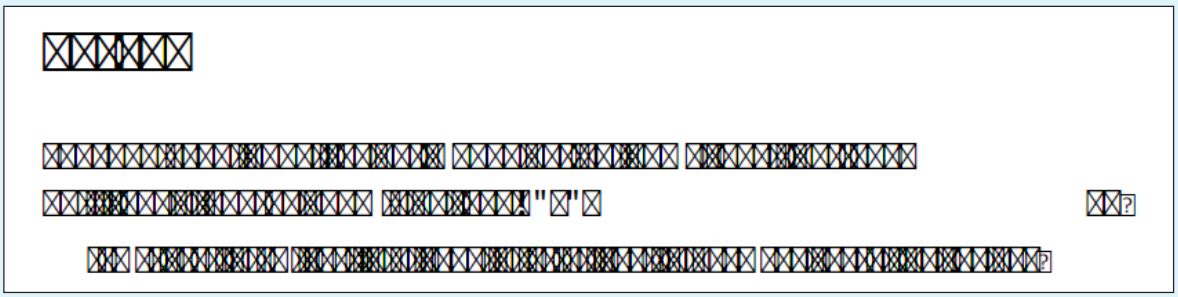


Obsah

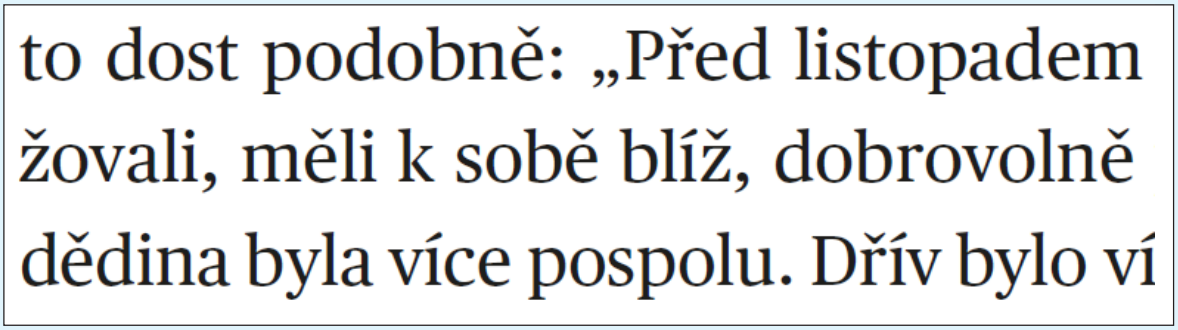
Znovuobjevení ohně: Jak nás mohou pasivní domy zbavit potřeby uhlí, ropy a jaderné energie do roku 2050 **11**

James Scott Brew, FCSI, AIA, CPHC, Director, Ebert & Baumann Consulting Enginee

Obrázek 10. Původní soubor PDF před konverzí

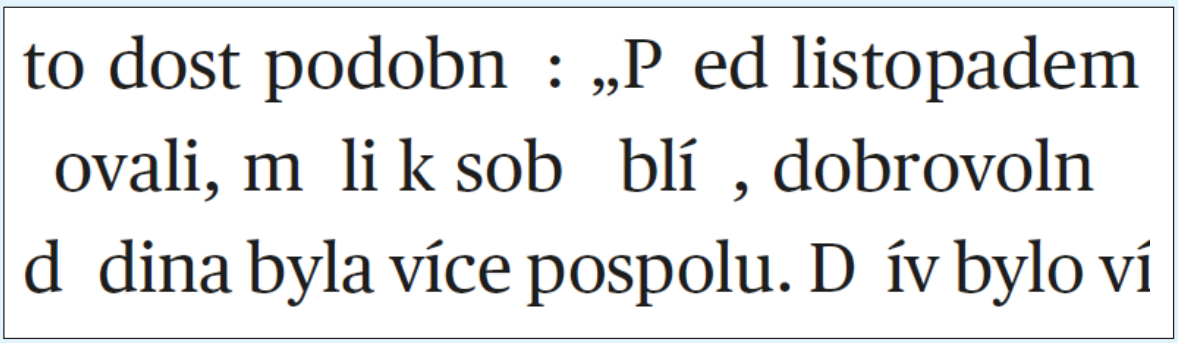


Obrázek 11. Výsledný soubor po převodu do PDF/A (nevalidní)



to dost podobně: „Před listopadem
žovali, měli k sobě blíž, dobrovolně
dědina byla více pospolu. Dřív bylo ví

Obrázek 12. Původní soubor PDF před konverzí



to dost podobn : „P ed listopadem
ovali, m li k sob blí , dobrovoln
d dina byla více pospolu. D ív bylo ví

Obrázek 13. Výsledný soubor po převodu do PDF/A (nevalidní)

Obrázek 14. Původní soubor PDF před konverzí

Obrázek 15. Výsledný soubor po převodu do PDF/A (nevalidní)

verze k zásadnější ztrátě informačního obsahu.
Provedené zkušební konverze mezi formáty PDF a PDF/A na reálných datech vnímáme jako jeden z řady kroků, které mohou pomoci správně nastavit formátovou politiku a ochranná opatření, která by měla v dlouhodobém horizontu minimalizovat rizika ztráty významné části písemného kulturního dědictví. Jako vhodné se do budoucna ukazuje otestování dalších konverzních nástrojů a jejich porovnání oproti využitému Adobe Acrobatu z hlediska úspěšnosti konverzí, validnosti nebo zachování signifikantních vlastností výsledných souborů. Nezbytným se jeví také podrobnější prozkoumání a hledání řešení pro jednotlivé typy chyb, bránících úspěšným převodům.

■ Článek bol recenzovaný

Poznámky:

[1] Článek vychází částečně z bakalářské práce Václava Jirouška obhájené na Ústavu knihovnictví a informační vědy Filozofické fakulty Univerzity Karlovy v roce 2018 (Jiroušek, 2018). Analýzy databáze ohlášených e-knih, vzorku e-publikací přijatých Národní knihovnou ČR a zkušební převody mezi formátem PDF a PDF/A byly realizovány na základě institucionální podpory dlouhodobého koncepčního rozvoje výzkumné organizace poskytované Ministerstvem kultury ČR.

[2] K signifikantním vlastnostem aktuálně, viz například Beňačková, Kočišová, Kopský, Mally a Ostráková (2020)

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.

BEŇAČKOVÁ, Miroslava, Pavlína KOČIŠOVÁ, Vojtěch KOPSKÝ, Frank Richard MALLY a Natalie OSTRÁKOVÁ, 2020. Signifikantní vlastnosti: příspěvek ke kolektivnímu nevědomí. ProInflow: Časopis pro informační vědy [online]. Brno: Katedra informačních studií a knihovnictví, Filozofická fakulta Masarykovy univerzity, 12(2), 21 – 44 [cit. 2022-05-17]. ISSN 1804-2406 (online). Dostupné z: https://doi.org/10.5817/ProIn2020-2-3

BJARNASON, Baldur, 2012. EPUB javascript security. Baldur Bjarnason [online]. 27 July 2012 [cit. 2022-05-17]. Dostupné z: https://www.baldurbjarnason.com/notes/epub-javascript-security/

BLÄSI, Christoph a Franz ROTHLAUF, 2013. On the Interoperability of eBook Formats [online]. Germany: Johannes Gutenberg-Universität Mainz, April 2013 [cit. cit. 2022-05-17]. Dostupné z: https://www.booksellers.org.uk/BookSellers/BizFormFiles/936121cb-a426-46da-b9aa-db8bd285d21e.pdf

BRITISH LIBRARY, 2021. File Formats Assessments [online]. Digital Preservation Coalition, Last modified on 29 June 2021 [cit. 2022-05-17]. Dostupné z: http://wiki.dpconline.org/index.php?title=File_Formats_Assessments

CUBR, Ladislav, 2017. Autenticita a digitální informace. Praha, 227 s. Dostupné také z: https://is.cuni.cz/webapps/zzp/detail/105596/. Disertační práce. Univerzita Karlova, Filozofická fakulta, Ústav informačních studií a knihovnictví. Vedoucí práce Prof. RNDr. Jiří Ivánek, CSc.

CUBR, Ladislav, 2010. Dlouhodobá ochrana digitálních dokumentů. Praha: Národní knihovna ČR, 154 s. ISBN 978-80-7050-588-5.

DAY, Michael, Peter MAY, Paul WHEATLEY, Simon WHIBLEY a Johan VAN DER KNIJFF, 2019. EPUB Format Preservation Assessment. In: Digital Preservation Coalition Wiki [online]. Version 1.4. British Library Digital Preservatin Team, 14 s. [cit. 2022-05-16]. Dostupné z: https://wiki.dpconline.org/images/7/73/EPUB_Assessment_v1.4a.pdf

DAY, Michael, David RUSSO a Simon WHIBLEY, 2019. Mobipocket Format Preservation Assessment. In: Digital Preservation Coalition Wiki [online]. Version 1.1. British Library Digital Preservatin Team, 12 s. [cit. 2022-05-16]. Dostupné z: https://wiki.dpconline.org/images/5/5b/Mobipocket_Assessment_v1.1a.pdf

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

HUTAŘ, Jan, Andrea MIRANDA, Eliška PAVLÁSKOVÁ, Zdeněk VAŠEK a Zdeněk HRUŠKA, 2018. Metodika logické ochrany digitálních dat. Knihovna AV ČR, v. v. i. Schválená metodika DG16PO2RO44. Dostupné z: http://hdl.handle.net/11104/0282107

JEŘÁBEK, Antonín, 2011. ISBN a online publikace. Ikaros [online]. 15(1) [cit. 2022-05-16]. ISSN 1212-5075. Dostupné z: http://ikaros.cz/node/13593

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. 2022-05-16]. 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. 2022-05-16]. Dostupné z: https://talkingpdf.org/is-pdf-an-open-standard/

KNIJFF, Johan van der, 2012. EPUB for archival preservation [online]. Version for external distribution. National Library of the Netherlands [cit. 2022-05-16]. Dostupné z: http://openpreservation.org/system/files/epubForArchivalPreservation20072012ExternalDistribution.pdf

KNIJFF, Johan van der, 2013. EPUB for archival preservation: an up-

date. Open Preservation Foundation [online]. The Open Preservation Foundation [cit. 2022-05-16]. Dostupné z: http://openpreservation.org/blog/2013/05/23/epub-archival-preservation-update/

KNIJFF, Johan van der, 2015. Policy-based assessment of EPUB with Epubcheck. KB RESEARCH: Research at the National Library of the Netherlands [online]. Koninlijke Bibliotheek [cit. 2022-05-16]. Dostupné z: http://blog.kbresearch.nl/2015/03/13/policy-based-assessment-of-epub-with-epubcheck/

KAUR, Kirnn, Stefan HEIN, Sabine SCHRIMPF a Manuela HOLZMAYER, 2014. Report on DRM preservation [online]. APARSEN, 85 s. [cit. 2022-05-16]. Dostupné z: http://www.alliancepermanentaccess.org/wp-content/uploads/sites/7/downloads/2014/06/APARSEN-REP-D31_1-01-1_4_incURN.pdf

KIRCHHOFF, Amy a Sheila MORRISSEY, 2014. Preserving eBooks. In: DPC Technology Watch Report [online]. Great Britain: Digital Preservation Coalition, s. 1-31 [cit. 2022-05-16]. 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. 2022-05-16]. Dostupné z: https://ipres2017.jp/wp-content/uploads/15Marco-Klindt.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. 2022-05-15]. 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. 2022-05-17]. Dostupné z: https://wiki.dpconline.org/images/2/22/PDFA_Assessment_v1.0.pdf

LINDLAR, Michelle a Yvonne TUNNAT, 2017. How Valid is your Validation? A Closer Look Behind the Curtain of JHOVE. International Journal of Digital Curation [online]. 12(2), 268 – 298 [cit. 2022-05-15]. ISSN 1746-8256. Dostupné z: http://dx.doi.org/10.2218/ijdc.v12i2.578

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. 2022-05-15]. Dostupné z: https://ipres2017.jp/wp-content/uploads/35.pdf

LIBRARY OF CONGRESS, 2021. Recommended Formats Statement [online]. Washington, D.C.: Library of Congress, [2021] [cit. 2022-05-15]. Dostupné z: https://www.loc.gov/preservation/resources/rfs/index.html

OSTRÁKOVÁ, Natalie a Vojtěch KOPSKÝ, 2020. Posuzování souborových formátů z hlediska dlouhodobého uchovávání a návrh metodiky pro Národní knihovnu České republiky. Knihovna: knihovnická revue [online]. 31(2), 85-105 [cit. 2022-05-16]. ISSN 1801-3252. Dostupné z: https://knihovnarevue.nkp.cz/archiv/2020-2/recenzovane-prispevky/posuzovani-souborovych-formatu-z-hlediska-dlouhodobeho-uchovavani-a-navrh-metodiky-pro-narodni-knihovnu-ceske-republiky

PENNOCK, Maureen, Paul WHEATLEY a Peter MAY, 2014. Sustainability Assessments at the British Library: Formats, Frameworks, & Findings. In: Proceedings of the 11th International Conference on Digital Preservation, iPRES 2014, Melbourne, Australia, October 6 – 10, 2014 [online]. s. 141 – 149 [cit. 2022-05-16]. Dostupné z: https://fedora.phaidra.univie.ac.at/fedora/get/o:378110/bdef:-Content/get

SCHROEDER, Stan, 2022. Amazon's Kindle will finally support epub files: But (of course there's a but)... Mashable [online]. New York: Mashable, ©2021, May 3, 2022 [cit. 2022-05-16]. Dostupné z: https://mashable.com/article/amazon-kindle-epub

SVOBODA, Tomáš, Marie BALÍKOVÁ, Jaroslav KVASNICA, et al., 2015. Elektronické publikace v Národní knihovně ČR [online]. 1. Praha: Národní knihovna České republiky, 93 s. [cit. 2022-05-15]. ISBN 978-80-7050-654-7. Dostupné z: https://drive.google.com/file/d/0B46gpfbHV70tR28wSzYyUmxqOHc/view

TUNNAT, Yvonne, 2017. JHOVE – the one and only PDF validator. Open Preservation Foundation [online]. Open Preservation Foundation, 19th Dec 2017 [cit. 2022-05-15]. Dostupné z: http://openpreservation.org/blog/2017/12/19/jhove-the-one-and-only-pdf-validator/

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. 2022-05-17]. Dostupné z: https://wiki.dpconline.org/images/f/ff/PDF_Assessment_v1.5.pdf

WHIBLEY, Simon, Peter MAY a Paul WHEATLEY, 2017. EBook Summary Assessment. In: Digital Preservation Coalition Wiki [online]. British Library Digital Preservatin Team, s. 1 – 18 [cit. 2022-05-15]. Dostupné z: http://wiki.dpconline.org/images/0/04/EBook_Summary_Assessment_v1.0.pdf

WOLF, Julia, 2010. OMG WTF PDF: [PDF Ambiguity and Obfuscation] [online]. [cit. 2022-05-15]. Dostupné z: https://www.troopers.de/media/filer_public/13/51/135142fe-0768-4b53-92a3-fd2f-c7248ed2/tr11_wolf_omg_pdf.pdf