NACRX001KUVN
prvotní identifikátor

Dne: 16.09.2014 Naše značka: NA- 3436-4/12-2014 Vyřizuje/tel.: Stodůlka / 974 847 343

Vaše značka:

ZPRÁVA ZE ZAHRANIČNÍ PRACOVNÍ CESTY

MÍSTO: Hamburg, SRN

ÚČEL CESTY: Účast na worskshopu „Preserving PDF: identify, validate, repair“

ÚČASTNÍCI CESTY: Mgr. Zbyšek Stodůlka

ZPRÁVU PODÁVÁ: Stodůlka

NAVŠTÍVENÉ INSTITUCE: ZBW - Deutsche Zentralbibliothek für Wirtschaftswissenschaften

TRVÁNÍ CESTY: 31. 8. – 3. 9. 2014

DATUM VYHOTOVENÍ: 16. 9. 2014

SCHVALUJE: PhDr. Eva Drašarová, CSc., ředitelka Národního archivu

podepsáno elektronicky

TEXT ZPRÁVY ZE ZAHRANIČNÍ PRACOVNÍ CESTY

ČÁST VŠEOBECNÁ:

odjezd Praha 31. 8. 2014 v 12:29
pohraniční přechodový bod [CZ/D]: Schöna(Gr) 14:09
příjezd Hamburg 31. 8. 2014 v 19:35

odjezd Hamburg 3. 9. 2014 v 8:22
pohraniční přechodový bod [D/CZ]: Schöna Gr. 13:45
příjezd Praha 3. 9. 2014 v 15:27

ČÁST ODBORNÁ:

Akce pořádané sdružením Open Planet Foundation (dále též OPF) slouží ke společnému setkání uživatelů (zástupců paměťových institucí) a vývojářů, k vzájemné výměně zkušeností a představ vycházejících z rozdílné zkušenosti obou prostředí. Součástí těchto setkání jsou také praktické výstupy vzešlé ze skupinové práce, buď v podobě volně přístupných aplikací nebo návrzích metodik/manuálů, které OPF, paměťové instituce a vývojáři mohou využít ve své práci.

Mezi přítomnými byli mj. zástupci Národního archivu Nizozemí, Dánska či Švédska, Francouzské národní knihovny, Bavorské státní knihovny nebo konsorcia německých národních oborových knihoven GoPortis. Přítomni byli i zástupci původců, kteří se zabývají střednědobým ukládáním dokumentů v digitální podobě, jmenovitě nizozemského ministerstva spravedlnosti nebo univerzitních pracovišť z Velké Británie (Archeology Data Service, Imperial College atd.). Za vývojářskou komunitu participoval předseda sdružení PDF Association Olaf Drümmer, dále zástupce společnosti Preservica (Tesella) a nezávislý vývojář z Finska.

V úvodní části **Ed Foley** z OPF shrnul problematiku stále rostoucího význam formátu PDF v paměťových institucích, ale také problémy, které nároky na multiplatformitu tohoto formátu institucím přinášejí.

Olaf Drümmer představil genezi vývoje formátu PDF zejména s ohledem na jeho ISO standardizaci a práce na nové verzi ISO 32000-2:2016. Zároveň poukázal na skutečnost, že ne všechny standardizace PDF byly úspěšně přijaty. Zatímco PDF/X (2001-2010) pro výměnu PDF tiskových výstupů a PDF/A (2005-2012) pro dlouhodobé uchovávání byly nejúspěšnější, PDF/E pro technickou dokumentaci se nedočkal téměř žádné odezvy, pomalu je přijímán PDF/VT (2010) pro variabilní data při tisku. Formát PDF/UA (2012) pro usnadnění přístupu snad v blízké budoucnosti čeká úspěch.

Dále se ve svém příspěvku věnoval vlastnostem formátu PDF/A a nárokům či problémům, které sebou využití tohoto formátu přináší a to i z historické perspektivy jejich vývoje v IT (přesná pravidla, nezávislost na vnějších zdrojích, nezávislost na zařízení a nástrojích, Unicode kódování textu, modely barev, omezení problematických vlastností v podobě šifrování, JavaScriptu, nevhodné syntaxe atd.).

Zabýval se podmínkami, za kterých je možné považovat formát PDF (nejen PDF/A) za vhodný formát pro střednědobé/dlouhodobé uložení: soulad se všemi pravidly ISO 19005 (PDF/A)/ISO 32000 (PDF

1.7), plný soulad se všemi pravidly odkazovaných specifikací a standardů (fonty, komprese obrázků, ICC profily atd.), dostatečné rozlišení obrázků, dobře skenovaná PDF obsahující bezchybný text pro OCR, obrazový obsah je přesně reprezentován, text je kódován jako text a může být mapován do Unicode pro extrakci, je možné využít strukturu obsahu (např. při migraci), vizuální zobrazení je vždy věrné originálu a 100% konzistentní.

Na závěr se věnoval nastínění problematických otázek, např. tomu, jaká pozornost se věnuje náležitostem formátů a není přitom věnována pozornost prohlížečům, které porušují standardy za účelem (téměř jakéhokoliv) úspěšného zobrazení, a kumulace z toho plynoucího rizika v čase.

Odpolední část byla věnována praktickému krátkému seznámení s identifikačními a validačními nástroji a se vzorovými sadami tak, aby je mohli využít všichni účastníci. Ti byli následně rozděleni do skupin, z nichž jedna se zabývala klasifikací závažnosti chyb validačních výstupů spojených s možností jejich opravy, druhá se věnovala vytvoření skriptu, který by poskytoval srovnání chyb při validaci pomocí Apache Preflight k jednotlivým produkčním nástrojům za účelem dalšího analytického vyhodnocení. Třetí skupina se věnovala definování požadavků na freewarový PDF/A konvertor. Dílčí výstupy byly průběžně posuzovány a diskutovány druhými skupinami.

Program druhého dne byl věnován problematice validace a validátorů formátu PDF/A. **Olaf Drümmer** poukázal, že validace PDF/A-2 se řídí 17 pravidly dle ISO 19005-2, 79 dle ISO 32000-1 a nepočetným množstvím specifikací, na které tyto pravidla odkazují. Problém je, že implementace často popisují struktury dat bez omezení, implementace specifikací závisí na vývojáři a ne vždy je zřejmé, co tím zamýšlel. Standardy se sice snaží nastolit normativy, ovšem jejich modalita vede k opaku, bez návodu, jak má být implementace správně. Zároveň paměťové instituce při využití svých nástrojů preferují nějaký výsledek než jen perfektní výsledek, s hlavním cílem zajistit stabilitu a použitelnost.

Dále se věnoval problematice validátorů, neboť jejich vytvoření je velmi náročné. K jejich zkoušení by měly napomoci testovací sady (k PDF/A-1 je to Isartor Test Suite), ale je to obecně nevděčná práce s problematickými výsledky. Komerční software se zaměřuje na mainstream a specifické požadavky paměťových institucí stojí na pokraji zájmu. Současné validátory kontrolují proti definovaným pravidlům v ISO 19005, nikoli proti odkazovaným pravidlům. Dodavatelé kooperují na sladění výstupů, aby usnadnili vzájemnou komparaci.

Při validaci PDF je podle Olafa Drümmera nutné zhodnotit rizika, která mají okamžitý dopad nebo která mohou mít dopad v budoucnu a to s ohledem na cíl, zda zobrazit soubor nebo zachovat všechny jeho vlastnosti (linky, textová extrakce, komprese obrázků atd.). Na co se pozornost při validaci málokdy soustředí je kódování v Unicode, struktura obsahu v tagovaném PDF, metadata, kódování barev, kvalita obrázků atd.). Zároveň se opomíjí kontrola kompletnosti obsahu nebo správné verze dokumentu. Kontrola syntaxe je přitom nenáročná a její dopad může být v budoucnu značný.

Zatímco ochota hradit validátory je nízká, není vzhledem ke komplexnosti a zároveň úzkému zájmu vůle k vytváření ze strany vývojářů svobodného software. Situaci se snaží zvrátit projekt PREFORMA hrazený z prostředků EU pro léta 2014-2017 (za účasti OPF a PDF Association).

Zatímco pracovní skupiny pokračovaly dále na řešení stanovených problémů, představily zástupkyně knihovnického konsorcia GoPortis praktické zkušenosti s přijímáním a ukládáním dokumentů. Sdílí

společnou základnu v podobě LTP systému ExLibris. **Yvonne Friese** z Národní ekonomické knihovny v Hamburgu přiblížila současný stav a budoucí plány zejména s ohledem na identifikaci a validaci formátů, jak ji provádí její instituce. K identifikaci používají DROID, ale také řadu vlastních vyvinutých jednoduchých nástrojů (kontrola tagu %PDF, verze, zda XMP, otevření, šifrování). Validace probíhá skrze JHOVE, s cílem v budoucnu využít PDFBox nebo PdfTron a KOST-Tools. Současnou strategií v případě, že soubor ve formátu PDF nelze opravit (např. obsahuje šifrování), je uložit jej. Opravy jsou prováděny vlastním iText pluginem, ale zvažován je i Ghostscript nebo Acrobat. K porovnání výsledku vyvinula referující vlastní program PdfTwinTest (porovnává řádky, iText vždy stejný, Ghostscript a Acrobat se liší často kvůli nahrazení fontů), ráda by časem zařadila OPF vyvíjený nástroj Matchbox. Upozornila, že opravování dokumentů ve formátu PDF může být zrádné a uvítala by jednodušší a lépe popsané nástroje k validaci.

Michelle Lindlar z Národní knihovny pro vědu a technologie v Hannoveru představila pohled na infrastrukturu GoPortis z institucionálního hlediska. Musí se vyrovnat s tím, že ze strany producenta dojde k ověření obsahu, ale již jej nezajímá, jaké problémy jím vytvořený dokument skýtá. Knihovna má omezené možnosti normalizace, neboť producent již nezkontroluje obsahovou správnost obsahu po jejím provedení. Při migraci sice existuje auditní stopa, ale vzrůstají požadavky na úložiště (originál ponechává). Při emulaci sice odpadají výše uvedené problémy, ale zároveň je třeba řešit problémy s právy duševního vlastnictví.

Referující pak představila testovací vzorek ze svého pracoviště, v podobě 6.351 PDF souborů elektronicky odevzdaných disertačních prací, které analyzovala programem pdfinfo, zejména s ohledem na obsah polí Creator/Producer za účelem rozlišení nejčastějších chyb a možnostech jejich eliminace již na straně producenta. U pole Creator bylo rozlišeno 888 nástrojů, nejčastěji PScript verze 5.2 (cca 750), LaTeX (cca 650), PScript veze 5.2.2 (cca 430), TeX (400), Acrobat PDF Maker for Word (cca 230) atd. U označení Producer se vyskytuje 492 různých nástrojů, v první desítce je devětkrát zastoupen Acrobat Destiller (první je verze 5.0 s 780 záznamy), tuto hegemonii narušil na šestém místě Dvipdfm (210 výskytů). Závěrem zdůraznila, že migrace je preferovanou ukládací strategií a to zejména s ohledem na průkaznost změn dokumentu (při zachování originálu). Zároveň ale naznačila, že naprostá důvěra v PDF metadata není na místě.

Seminář byl ukončen závěrečnou prezentací dosažených výsledků skupinové práce a diskusí. Bylo zřejmé, že přítomní zástupci paměťových institucí by uvítali více standardizovaných nástrojů pro hromadné zpracování dokumentů ve formátu PDF resp. PDF/A a zároveň také více metodických materiálů zohledňující kategorizaci rizik a možných kroků při nápravě u dlouhodobého ukládání v těchto formátech. Z neformální diskuse vyplynulo, že všichni zástupci archivů řeší nastavení svých workflow při příjmu jak s ohledem na dosažení svých ukládacích strategií, tak s ohledem na jejich efektivitu (např. téměř nereálné vizuální porovnávání výsledků migrace). Vzhledem k povaze setkání bylo možné s kolegy hovořit o problémech, se kterými se vyrovnávají ve svých institucích, což nemálo přispělo ke kooperativnějšímu prostředí než na jiných formálnějších setkáních k problematice dlouhodobého ukládání, kde se prezentované výsledky mnohdy rozcházejí s vlastní realitou.