univerzitná knižnica
v bratislave
Univerzitná knižnica v Bratislave | Michalská 1 | 814 17 Bratislava

Preforma Experience Workshop – Improving long-term digital preservation

Dňa 23. 11. 2016 sme sa 4 pracovníci Centrálneho dátového archívu zúčastnili medzinárodnej konferencie v Berlíne, ktorá sa venovala dlhodobému uchovávaniu. Išlo aj podľa názvu konferencie o workshop, výmenu skúseností medzi inštitúciami, ktoré sú súčasťou konzorcia PREFORMA. Cieľom bolo demonštrovať zhodu súborových formátov vyvinutých v rámci projektu za účasti pamäťových inštitúcií mimo konzorcia PREFORMA pri skúšaní, využívaní a ďalšom rozvoji softvéru a zdieľanie skúseností získaných z pamäťových inštitúcií pracujúcich s vývojármi na základe zmlúv o poskytovaní služieb v oblasti výskumu a vývoja. CDA sa stalo spolupracujúcou inštitúciou v roku 2016. Radi by sme v budúcnosti testovali v spolupráci s PREFORMA konzorciom nástroje, ktoré vyvíjajú na validáciu pdf, tiff a video formátov.

Borje Justrell – projektový koordinátor PREFORMA, Swedish National Archives predstavil projekt: PREFORMA znamená PREservation FORMAts pre archívy a ide o projekt spolufinancovaný Európskou úniou. Cieľom projektu je, aby mali pamäťové inštitúcie plnú kontrolu nad procesmi testov zhody súborov na vytváranie, migráciu a ingest do archívov, na overenie zhody so štandardmi formátov (validátory). Výsledkom projektu by mali byť voľne dostupné nástroje. Hlavnou témou jeho prezentácie bola problematika formátov. Hovoril, že pamäťové inštitúcie čelia narastajúcemu transferu elektronických dokumentov a iných digitálnych obsahov pre dlhodobé uchovávanie. Tento obsah sa ukladá v špecifických formátoch pre dokumenty, obrázky, zvuk, video, atď., a tieto súbory sú často produkované softvérmi od rôznych vendorov. Uviedol, že dokonca, ak aj sú súbory v štandardných formátoch, nie je zaručená správna implementácia štandardov. Je to z toho dôvodu, že softwary, ktoré sa používajú na produkciu elektronických súborov, nie sú pod kontrolou inštitúcií, ktoré ich produkujú a ani pod pamäťovými inštitúciami. Testy zhody sú vykonávané pamäťovými inštitúciami, ale nie sú celkom spoľahlivé, a odlišné softwary pre

testovanie môžu skončiť s rozličnými výsledkami. To predstavuje problémy dlhodobého uchovávania.

Ďalej hovoril o tom, že PREFORMA sleduje open source prístup, s cieľom vytvoriť udržateľný výskum a vývoj, zahŕňajúci širokú škálu prispievateľov a užívateľov z rôznych skupín, a uviedol výhodu open source prístupu, ktorý zabezpečuje dlhodobú dostupnosť softvéru. Tieto softvéry pre PREFORMU vyvíjajú títo dodávatelia: 1. PDF validátor vyvíja VeraPDF – validátor pre PDF/A, 2. TIFF validátor vyvíja Easy Innova a 3. Validátor pre audiovizuálne súbory (MKV, FFV1, LPCM) vyvíja MediaArea, títo v druhej časti konferencie predstavili validátory pre uvedené formáty. Na záver uviedol budúce podujatia, ktoré sa budú konať v rámci PREFORMA konzorcia – v marci 2017 konferencia v Padove a záverečná konferencia je naplánovaná na jeseň 2017 v Estónskom Talline, kde budú predstavené výsledky projektu.

Ďalším prezentujúcim bol Hannes Kulovits z Austrian State Archives, ktorý hovoril o dlhodobom uchovávaní v Rakúsku. Zaujímavé bolo, že Rakúsko v roku 1999 prijalo federálny akt bezpečného uchovávania, archivovania a používania archívnych záznamov vo vláde. Podľa neho je kľúčovou rolou živého archívu, že vláda musí ponúknuť aktuálne záznamy archívu (13 ministerstiev plus podriadené agentúry) a preniesť ich do archívu po dobu 10 rokov. V ich archíve uchovávajú „Made-digital“ dáta a „Born-digital“ dáta (elektronické záznamy >= 2003). Od roku 2004 boli všetky záznamy „Born-digital“, používajú Centrálny elektronický manažment systém záznamov, a v roku 2013 mal 11000 používateľov a 18 TB dát.

Uviedol aj niekoľko výziev, ktorým musia čeliť, zaujímavé to bolo z toho hľadiska, že v mnohých oblastiach je aj na Slovensku podobná, resp. rovnaká situácia: ministerstvá sú uvoľnené z dodržiavania smerníc, nejednotná politika záznamov naprieč ministerstvami, žiadne formátové obmedzenia, softvérové poruchy a pod. Vo veľkej miere sa venoval formátom a hlavne nejednotnosti a veľkej diferencii formátov, s ktorou sa stretávame aj v CDA. Preto by som uviedla výrok, ktorý mal vo svojej prezentácii “We do not currently have specifications for these older file formats.” “It is likely that those employees who had significant knowledge of these formats are no longer with Microsoft.”(Tony Hey, Corporate Vice President Microsoft Research).

Uviedol niekoľko rizikových faktorov pri formátoch: množstvo voľne dostupných nástrojov, nedostupnosť dokumentácie, kvalita dokumentácie, štandardizácia, možnosti identifikácie, možnosti validácie, práva duševného vlastníctva, cena za licencie. Na záver prezentácie ukázal ukážky z ich archívneho systému a predstavil use case jedného problému, kde hovoril o prípade pdf, ktorý sa vykresľoval rozdielne na rôznych počítačoch, pdf bol digitálne podpísaný, pdf súbory boli úplne identické, skontrolované pomocou kontrolných súčtov checksum. Problém vznikol, keď systém zistil, že nebolo vložené písmo a počas zobrazovania aplikácia načítala fonty z operačného systému a pokiaľ font neexistoval, systém ho nahradil podobným a štruktúra dokumentu sa mohla zmeniť. Na validáciu požívajú tieto nástroje: JHove, FITS, DROID, veraPDF, DPF M.

Z tohto podľa neho vyplýva, že kľúčom k úspechu digitálneho uchovávania by malo byť, aby pamäťové a fondové inštitúcie vedeli, aké dáta majú, mali by si byť isté, že digitálne súbory sú také, aké by mali byť a treba vedieť aké opatrenia prijať a ako.

S ďalšou prezentáciou sa predstavil Matthias Priem z firmy VIAA (Vlaams Instituut voor Archivering), v ktorej sa zaoberajú digitalizáciou, archiváciou a diseminációu, pracujú pre viac ako 100 organizácii z rôznych oblastí. Uchovávajú aj videá, mesačne im do archívu

pribudne približne 3000 položiek „born-digital“ dát, čo predstavuje cca 60 TB. Majú vlastný systém, dátové centrum majú v Belgicku na 3 geograficky oddelených miestach. Zaujímavé v tejto prezentácii bolo najmä to, že výstupy z diseminácie úspešne využívajú vo vzdelávaní použitím prezentačného portálu. V závere prezentácie sa tiež venoval formátom, hovoril, že najlepšie je zvoliť hneď na začiatku jednotný formát, ak je to možné, pre všetky dáta, a urobiť testy zhody predtým, než začnú ostré vklady. Hovoril, že zložitejšie to je s born-digital formátmi, sú omnoho rozmanitejšie, od mnohých dodávateľov, a vplýva na ne rýchly technologický vplyv. Skutočnou výzvou pri uchovávaní sú teda podľa neho born-digital dáta a uchovávanie nových formátov. Pre CDA bolo zaujímavé, keď hovoril o nastavovaní uchovávania pre mediálnu oblasť, aplikujú FFv1 (FF video codec 1), pokúšajú sa konvertovať jp2k do FFv1 bez dátovej straty. Predbežné výsledky pokusov: dekódovanie na mp4 použitím ffmpeg sa javí 3-4 krát rýchlejšie, keď sa začínalo z FFv1, Jpeg 2000 vs FFv1 – je treba o 10% menej miesta, bezstratové transkódovanie sa javí ako možnosť. Niektoré ďalšie veci sú v procese testovania.

Uchovávaním videa sa zaoberal aj Erwin Verbruggen z Netherlands Institute for Sound and Vision. Tieto spoločnosti sa okrem iného zaoberajú aj uchovaním digitálneho videa vo formáte MXF (Media eXchcage Format). Navyše spoločnosť VIAA kooperuje na štandardizácii formátu FFv1, ktorý sa javí ako jeden z potenciálnych formátov pre archiváciu videa v CDA. Spomenutá bola aj iniciatíva IIIF (International Image Interoperability Framework).

Erwin Verbruggen predstavil implementáciu OAIS modelu v ich archíve, na ktorom je založený aj archív CDA, a uviedol budúci plán certifikácie DSA (Data Seal of Approval) ich archívu.

V ďalšom bloku prezentácii bol nosnou témou výber stratégie pri digitálnom uchovaní dát (prezentácia p. Lemmensa z PACKED) a prehľad súčasných štandardov pre jednotlivé formáty (prezentácia p. Gebera a p.Yuosefiho z Riksarkivet).

Hovorilo sa napríklad aj o migrácii objektov do niektorého formátu. Dôležité pri formátovej migrácii je zhodnotiť riziká, ktoré pri tom hrozia. Dôležité je používať otvorené formáty, ku ktorým existujú špecifikácie, a ku ktorým sa bude možné dostať aj o niekoľko

rokov. Neznamená to však to, že keď vieme súbor otvoriť teraz, budeme to vedieť aj neskôr, potom môžu nastať aj také situácie, že súbor otvoríme, ale v ňom budú len prázdne štvorčeky bez akejkoľvek výpovednej hodnoty. Preto je veľmi dôležité držať sa špecifikácii formátov. Proprietárne formáty môžu byť špecifikované, a podrobný popis špecifikácie môže vlastniť len firma, ktorá formát používa, a už nastáva problém. Preto sa PREFORMA snaží vytvoriť nástroje na validáciu nižšie spomínaných formátov, a hlavne, aby k nim existoval open-source prístup a špecifikácie.

V záverečnej časti workshopu boli prakticky odprezentované nástroje na validáciu formátov: VeraPDF – na validáciu PDF/A (http://verapdf.org/software/), DPF manager (http://dpfmanager.org/) na validáciu TIFF, MediaConch (https://mediaarea.net/MediaConch) na validáciu súborov MKV s kodekom FFv1.

V Bratislave, dňa 11.01.2017

Mgr. Lucia Kelemenová