Zpráva ze služební cesty do USA – Washington

Účel cesty: Digitalizace v Kongresové knihovně (LOC) a Národním archivu (NARA)

Termín cesty: 15. 10. - 21. 10. 2011

Účastníci cesty: PhDr. Karel Koucký, PhDr. et Ing. Milan Vojáček, PhD., Národní archiv; Mgr. Tomáš Dvořák, Archiv hlavního města Prahy

Zprávu podává:

PhDr. Karel Koucký, Národní archiv, v.r.

Datum vyhotovení: 22. 12. 2012

Podpis ředitelky archivu:

Organizační údaje o služební cestě

Odjezd z Prahy: dne 15. 10. 2011 v 7.00 hod. letecky

Příjezd do Washingtonu: dne 15. 10. 2011 ve 14.30 hod.

Odjezd z Washingtonu: dne 20. 10. 2011 v 18.00 hod. letecky

Příjezd do Prahy: dne 21. 10. 2011 v 11.30 hod.

Ubytování: v hotelu ve Washingtonu.

Vyúčtování cesty: provedeno v NA v předepsané lhůtě po návratu.

Průběh pobytu: kromě doby určené na přepravu probíhaly konzultace na dohodnutých pracovištích částečně podle předem dohodnutého programu, některé části programu byly upraveny na místě.

Průběh a výsledky jednání

Cesta byla realizována v rámci řešení společného výzkumného projektu Národního archivu a Státního oblastního archivu v Praze „Zajištění ochrany archivních dokumentů důležitých pro potřeby státu“ financovaného z prostředků Programu bezpečnostního výzkumu České republiky v letech 2010 - 2015 (BV II/2 – VS) (identifikační kód projektu VG20112014054). Cílem pracovní cesty bylo získat vyčerpávající informace o projektech, programech a iniciativách Kongresové knihovny a Národního archivu v oblasti digitalizace využitelné pro potřeby výzkumu strategií vytváření bezpečnostních kopií archiválií.

Národní archiv (National Archives and Records Administration, NARA)

Jednání proběhla na několika specializovaných pracovištích v detašovaném umístění archivu v College Park.

ERA Program Office

David C. Lake představil projekt ERA, který se zabývá dlouhodobým uchováváním digitálních dokumentů. Počátek projektu digitálního archivu se datuje od r. 1999/2000. Komerčním parterem se v počátku stala společnost Lockheed Martin Corporation. Příjem a

zpracování dokumentů probíhaly manuálním způsobem s cílem automatizovat tyto procesy. V září 2011 byl tento projekt dokončen. Data se do současnosti předávala převážně na médiích (CD, DVD). Dílčím výsledkem projektu ERA je stejnojmenný softwarový produkt. Jedná se o informační systém vyvinutý společností Lockheed Martin ve spolupráci s dalšími subdodavateli (např. Tessela), který se skládá např. z datového úložiště, native XML databáze fy. MarkLogic, Oracle databáze a dalších programů. Z funkčního hlediska je digitální archiv postaven na modelu OAIS. Informace jsou do archivu předávány v SIP balíčcích ve formátu ZIP, u kterého se kontroluje hash celého balíčku. Během příjmu do archivu probíhá indexace klíčových slov. Dále se používá program JHOVE pro ověření souborového formátu dokumentů a registr PRONOM jako informační systém o těchto formátech. Ačkoli nejsou vstupní formáty legislativně stanoveny, pro projekt ERA se používají formáty PDF, TIFF, JPEG a patrně i další. Informace o dokumentech jsou uchovávána v XML metadatovém modelu PREMIS. Po příjmu se data ukládají na datové úložiště. To má v tuto chvíli kapacitu 130 TB. Data, která budou po dlouho dobu nepřístupná, nejsou z důvodu finančních úspor ukládána na disková pole s RAID jako ostatní. Na r. 2012 je plánován nákup Hierarchical Storage Management určený pro řízení ukládání dat. Dále disponují SW framework pro konverzi dokumentů a metadat.

Zodpovězeny byly některé připravené otázky. V péči o digitální (digital-born) a digitalizované archiválie zde není rozdíl. Digitalizované dokumenty jsou uloženy v digitálním archivu ERA, přičemž v jejich metadatech je uvedeno, že se jedná o digitalizované archiválie. Do doby, kdy ještě není zcela dokončen příjem do archivu, se data na straně původce neničí. Důležitá jsou data a nikoli médium, na kterém jsou data předávána. Na otázku ke ztrátě dat bylo odpovězeno, že data se ztrácí neustále, ale díky zálohám a technologii RAID 6 se daří data znovu obnovovat. Drtivou většinou dokumentů uložených v ERA jsou běžné dokumenty, databází a dokumentů GIS je jen malé procento. V otázce datové standardizace balíčku SIP jsou pracovníci NARA na začátku. Stanoveny jsou jen povinné údaje, nikoli datové schéma pro metadata. K této standardizaci však dojde později. Digitální archiv se nachází mimo hlavní budovu NARA (na blíže nespecifikovaném místě ve státě West Virginia). K dotazům na projektové řízení a nákladnost projektu bylo řečeno, že do projektu je zapojeno mnoho lidí (na straně NARA jde o 25 pracovníků), rozdělených podle úkolů na vývojáře, testery, systémové inženýry apod. Celý projekt ERA stál již 457 miliónů dolarů (vývoj digitální archivu, technické vybavení, projektový management, náklady na údržbu a operační náklady). Novým partnerem NARA je v této oblasti spol. IBM, které se bude po dobu 10 let starat o

provoz a údržbu digitálního archivu. Tyto náklady se pohybují mezi 25-30 miliónů dolarů ročně.

Diskuze s Davidem R. Kepley z téhož oddělení byla směřována k obecným koncepčním problémům dlouhodobé archivace digitálních dokumentů (digital-born, digitální reprodukce). Z projektového hlediska byly vysvětleny fáze konceptuálního rámce digitální archivace, strategie plánování uchovávání a zpřístupnění a další úkony. Nastíněn byl problém s obecnou definicí dokumentu. Namísto toho byla jako základní jednotka uchopena intelektuální entita informačního modelu OAIS a na jejím podkladě sestaven balíček. Velmi se zde dbá na autenticitu tohoto balíčku, který je uchováván v původní podobě a zároveň migrován pro účely čitelnosti do jiných formátů, přičemž je neustále sledováno a zaznamenáváno nakládání s balíčkem. Dokumenty, které jsou v digitálním archivu, jsou v různých formátech (v počtu cca 100), přičemž některé z nich jsou rizikové (relační databáze, zapouzdřené dokumenty, výstupy GIS, e-maily s přílohami).

Otázky byly cíleny k základním oblastem problematiky. Na otázku, zda má cenu pokračovat v e-governementu s ohledem na finanční náklady uchování bylo odpovězeno příměrem ke shodné situaci z počátků archivnictví, kdy i pro analogové dokumenty byla potřeba budovat stavby a dodávat vybavení, což se neobešly bez nákladů. Na otázku, zda bude možné dokumenty v různých formátech interpretovat donekonečna, bylo sděleno, že patrně nikoli. Budovat technologické muzeum je nemožné. S emulací, která nakládá vždy s originálem, se experimentuje v Nizozemí, ale pravděpodobně i v tomto případě se nejedná o správnou cestu. Je zřejmé, že bude nutné dříve či později dokument konvertovat do jiného formátu, ve kterém se uchovají jeho tzv. podstatné vlastnosti (significant properties; sestávají ze vzhledu, struktury, obsahu a chování) a proces migrace bude potřeba dokladovat. Jen tak je možné zaručit čitelnost a přitom udržet autenticitu v přijatelné rovnováze. I pro digitalizované archiválie je model ERA vhodný, neboť je nelogické oddělovat uchování digitálních a digitalizovaných dokumentů. Procesy uvnitř archivu jsou obdobné.

Applied Research Division

Richard Lopez, Rita Cacas a Mark Conrad představili oddělení aplikovaného výzkumu NARA, které spolupracuje na několika projektech samostatně nebo ve spolupráci s mnoha výzkumnými organizacemi (např. projekt ERA). Jsou zde vyvíjeny softwarové nástroje a databáze pro usnadnění zpracování digitálních dokumentů pro účely správy velkého množství dat, přičemž se jedná převážně o podpůrné prostředky určené digitálním archivům a knihovnám. Činnost oddělení byla demonstrována na příkladu řešení projektu na zpřístupnění

výsledků sčítání obyvatelstva z r. 1940. Všechny sčítací archy (cca 3,6 miliónů) byly digitalizovány a pomocí OCR rozpoznány předtištěné rubriky. V případě rozpoznávání rukopisných údajů dosáhli při aplikaci OCR zcela nepoužitelných údajů. Vyvinuli proto speciální algoritmus pro rozpoznávání ručně psaného textu, který vychází z dynamiky psaní (s odkazem na dosavadní výzkum v této oblasti). Jeho úspěšnost je kolem 70%. Pro účely indexace všech archů byly využity i lidské přepisy údajů, tedy celkově byl aplikován hybridní způsob rozpoznávání.

Information Services – Online Public Access, Research Services

Rebecca Warlow, M'Lisa Whitney a Mary Rephlo představily digitalizační strategii v NARA. Důraz je kladen na aktivní spolupráci se soukromými partnery ze strany genealogických společností (např. FamilySearch.org, Ancestry.com), kteří pokrývají náklady na digitalizaci. Skenují se jakékoli materiály, které obsahují jmenné údaje. Byl popsán postup digitalizace včetně důležité přípravy materiálu v konzervační dílně. Digitalizují se jak originály archiválií, tak mikrofilmy. Přináší to velká množství obrázků (cca 5-7 tisíc týdně), v současnosti má NARA cca 400 tisíc digitálních objektů. V oblasti metadat není situace uspokojivá. Cílem je kompletní zpřístupnění digitálních objektů v centrálním archivním katalogu, kde je v současnosti přístupných 126 tisíc digitálních objektů. Mezi NARA a konkrétním partnerem se uzavírá smlouva, v níž jsou dohodnuty obvyklé technické parametry snímání (formáty výstupů, rozlišení, barevná hloubka, jména souborů, formát metadat aj.), které z vnějšího pohledu naznačují spíše kvalitnější digitalizaci. Digitalizuje se výhradně v budově archivu. Pracovníci partnerů jsou vyškoleni v zacházení s archiváliemi zaměstnanci NARA, dále je jim poskytována podpora při digitalizaci; pracovníci NARA provádí kontrolu výstupů.

Otázky se týkaly vztahu mikrofilmování a digitalizace. Podle sdělení pracovnic NARA všechno, co se digitalizuje, se předtím mikrofilmuje pro bezpečnostní účely. Paní Whitney je ovšem toho názoru, že na mikrofilm nelze pohlížet jako na jediné bezpečné médium pro archivní kopie. Při výběru archiválií určených k digitalizaci přímo v NARA se uplatňují laické zájmy (široká veřejnost), stejně jako specifické zájmy užšího okruhu veřejnosti odborné. Digitalizační strategie NARA pro roky 2007-2016 je orientována především na zpřístupnění a je velmi obecná. V přístupu č. 3 a 4 se připouští bezpečnostní digitalizace v souladu se strategií změny formy archiválií v případě rizika poškození originálu. Zajímavý je náhled na digitální kopie, které jsou veřejným statkem a jsou poskytovány zdarma (na rozdíl např. od Velké Británie). Tyto kopie (nejčastěji vystavené na Internetu) může veřejnost libovolně použít i ke komerčním účelům.

Conservation Lab

Vedoucí laboratoře Mary Lynn Ritzenthaler představila pracovní činnosti zaměstnanců konzervátorské dílny.

Digitization Lab

V doprovodu pánů Jeffrey Reeda a Martina Jacobsona byla realizována prohlídka digitalizačních dílen NARA, ve kterých se skenují plošné archiválie. Menší archiválie skenují na deskových skenerech EPSON prostřednictvím programu SilverFast, s úpravami v Adobe Photoshop (v prostředí Mac). Černobílé fotografie se skenují na 400 ppi, v 8bitové barevné hloubce grayscale, do formátu TIFF. Časově trvá snímání jedné předlohy o velikosti A5 cca 60 s. Do obrázků se doplňují metadata ve formátu XMP. Na otázku k rozsahu editace technických metadat bylo sděleno, že se o této problematice pouze diskutuje, žádná zvláštní metadata se nevytvářejí. Pro velkoformátové předlohy jsou k dispozici Cruse Scanner (pro předlohy cca 1,5 x 3 m). Dotazy směřovaly k velikosti digitalizačních dílen. Pracuje v nich přibližně 50 pracovníků, přičemž 2/3 z nich jsou operátoři digitalizačních zařízení.

V dalším úseku byla představena péče o filmy (motion picture). Pro záchranu starých filmů se používá analogový převod. Převádí se kolem 50 kotoučů týdně. Tyto filmy se rovněž digitalizují prostřednictvím zařízení Spirit 4K. Náklady na zařízení pro konverzi se před 3 lety vyšplhaly na 900 tisíc dolarů, postprodukční studio pak na další 2 milióny dolarů. Jiný úsek se věnoval audiu/videu. Nachází se zde rozsáhlé technologické muzeum, které se skládá z různých přehrávačů původem od federálních původců. Tento úsek se zaměřuje na uchovávání a převod zvukových pásek a kazet a stejně tak i obdobných audiovizuálních médií (kromě klasických filmů). Datový formát JPEG2000 považují za nestabilní, data pro zpřístupnění ukládají do MPEG2. V mikrografickém úseku se digitalizuje pomocí zařízení Mekel Technology MACH IV, který snímá 270 políček za minutu. Digitalizuje se s rozlišením 600 ppi, v 8bitové hloubce grayscale. Toto zařízení je optimalizováno pro práci s pozitivem i negativem, preferován je pozitiv. Jako skenovací SW je používán nástroj QuantumScan, případně QuantumProcess. Nejstarší mikrofilm v NARA pochází z 50. let minulého století. Aktuálně jediným zařízením určeným pro převod obrazu na mikrofilm je technologii COM. Jedná se zařízení Zeutschel OP 500.

Na závěr prvního pracovního dne byla realizována prohlídka největšího digitalizačního pracoviště NARA s 11 velkoformátovými skenery Zeutschel formátu A1 a A2.

Druhý den v NARA byl celý věnován prohlídce hlavní veřejné budovy NARA situované v centru Washingtonu, ve které jsou umístěné expozice pro veřejnost. V průběhu prohlídky byly diskutovány přístupy NARA k propagaci archivnictví.

Kongresová knihovna (Library of Congress, LOC)

Úvodní jednání v LOC patřilo koordinátorovi výjezdu p. Stevovi Pugliovi, který představil LOC jako instituci s dlouhodobou tradicí v digitalizaci (počátky digitalizace spadají do pol. 90. letech 20. stol.). Knihovna se profiluje rozsáhlou domácí i zahraniční akviziční činností, k ukládání knih a dalších materiálů využívají řady úložných prostor ve městě i mimo něj (ve Washingtonu aktuálně uloženo cca 25 milionů jednotek, mimo město dalších 40 milionů). Knihovna vyvíjí standardy pro uchovávání digitálních objektů. V oblasti rozvoje své IT infrastruktury směřují spíše k vývoji jednotlivých SW nástrojů, nikoli komplexního informačního systému.

Preservation Directorate

Proces uchovávání popsala Mary Oey jako jednu z tradičních knihovních činností v komplexu péče o dokumenty. Při něm se snaží garantovat přístupnost obsahu dokumentů (content access), nikoli dokumentů jako takových (object access). Za tímto účelem se vytvářejí kopie ve formě mikrofilmů, digitálních snímků nebo xerokopií.

Diskutována byla otázka, zda LOC vnímá digitalizaci jako snímkování pro účely ochrany nebo zpřístupnění původních dokumentů. Paní Oey se vyjádřila v tom smyslu, že tyto funkce nelze od sebe dost dobře oddělit. Na druhou stranu si uvědomuje možný odlišný přístup např. u digitalizace středověkých rukopisů, která se provádí za účelem zpřístupnění. Naopak u audiovizuálních záznamů cílí jejich digitalizace k trvalému uchování. V LOC se nacházejí dva úseky, tzv. všeobecné (general) a zvláštní (special) sbírky, které se snímkují odděleně od sebe. Součástí knihovny je Národní audiovizuální konzervační středisko, které sídlí v městě Culpeper v Západní Virginii.

V knihovně probíhá mnoho digitalizačních projektů. U poškozených dokumentů se provádějí spektrální testy za účelem zvýšení čitelnosti zaznamenané informace. K otázce vyčíslení nákladů na uchování uložených dokumentů bylo sděleno, že nejlevnější je uchovat originál (prostřednictvím deacidifikace), více nákladů vyžaduje mikrofilmování a jako nejvíce nákladná se jeví digitalizace. Na druhou stranu je pro mnoho knihoven jednodušší přijímat mikrofilmy novin než jejich originály.

Preservation Research and Testing Division

Toto oddělení představila paní Cindy Connelly Ryan. V konzervační dílně pracuje 20 zaměstnanců, kteří převážně připravují dokumenty pro výstavy. Testovací laboratoře se věnují prověřování médií, na nichž jsou knihovní dokumenty uloženy a ve kterých se přejímají do knihovny (pásky, optické nosiče). Zejména se jedná o CD nosiče. Na testech těchto médií nespolupracují s technickými univerzitami, ale se standardizačním institutem. Životnost CD nosiče se zde udává na 30 let, CD se zlatou vrstvou mají tuto životnost delší. Neexistují však průměrná média, pro které by mohla tato průměrně odhadnutá životnost platit, vždy je nutné posuzovat konkrétní média. Ukázána byla zařízení pro provádění spektrální analýzy médií. Životnost moderních nosičů se obecně testuje simulováním podmínek stárnutí v podobě extrémních klimatických podmínek. Byla nám poskytnuta vydaná studie z r. 2005 o životnosti magnetických a optických médií (Predicting the Life Expectancy of Modern Tape and Optical Media).

Federal Agencies Digitization Guidelines Initiative

O projektu FADGI informoval Steve Puglia. Do projektu je zapojeno více než 22 organizací, které se věnují převážně technickým aspektům digitalizace. Vedle výzkumu kvality výstupů digitalizačních projektů by se rádi zabývali i autenticitou těchto výstupů. V rámci výzkumu kvality snímaného obrazu byly sestaveny kalibrační terčíky a vytvořen open-source program pro rozpoznávání těchto terčíků. Záměrem je celou tuto sadu standardizovat u ISO. Na otázku vztahu mirkofilmování a digitalizace a jejich preference bylo sděleno, že přístupy se u různých institucí liší a že se liší i v rámci LOC. Rozpočet na fungování FADGI v rámci LOC je 100 tisíc dolarů ročně.

Preservation Reformatting Division

Činnost tohoto oddělení představila vedoucí Adrija Henley společně se svými kolegy. V 70. letech se v LOC začalo s mikrofilmováním monografií a časopisů, později i novin. Mikrofilmuje se z důvodu zamezení poškozování dokumentů v průběhu jejich fyzického používání, mikrofilmové kopie se vytvářejí i pro účely zpřístupňování. Digitalizace je vnímána jako způsob zpřístupňování obsahu dokumentů. Jejich výzkum se soustřeďuje na vytěžování dat z mikrofilmů a vytváření standardů v oblasti kvality reprodukcí. Z mikrofilmových svitků lze mnohem snadněji vytvářet digitální snímky (rychleji a levněji), které budou k dispozici v příslušné studijní kvalitě. K této oblasti nám bylo předáno několik

odborných analýz. V závěru jednání jsme navštívili trezorový depot s bezpečnostními mikrofilmy.

Manuscript Division

V závěru pracovního dne nám bylo předvedeno, jakým způsobem probíhá v LOC digitalizace zvláštních sbírek. Digitalizuje se zde s využitím technologie Sinar Phase One, který ovšem není příliš vytěžován. Zajímavým zjištěním bylo, že v průběhu snímání není proměřováno osvětlení přístroje a tato světla nejsou ani pravidelně měněna. Dále se zde používá knižní skener i2S Suprascan, reprodukce se ukládají výhradně do formátu TIFF. Zvláštní technická metadata nevytvářejí. Další stroje i celkový přístup byl víceméně konvenční.

Office of Strategic Initiatives

Steve Puglia společně s kolegy poskytli informace k využívanému CMS iRODS (The Integrated Rule Oriented Data System) určenému pro práci s digitálními objekty a jejich evidenci. Dále se zmínili o projektu digitalizace novin. Podařilo se jim digitalizovat 1 bilión stránek, což představuje celkové množství 0,15% všech novinových stran, které jsou k dispozici v LOC. Hovořili o množství dat rovnajícího se 1 PB. Za tímto účelem používají pro ukládání obrázků formát JPEG2000 ve ztrátové kompresi. Dále zmínili, že zde není k dispozici centrální úložiště, ale že je využíváno decentralizované řešení s ohledem na decentralizovanou technickou infrastrukturu. V takových případech (bez centrálního řešení) se jim může stát, že některé on-line zdroje nebudou přístupné.

Collections & Services Directorate

Michael Neubert poreferoval o počátcích digitalizace v LOC. Prioritně zde byly od počátku skenovány předlohy, které byly zajímavé pro veřejnost (dokumenty z občanské války, jakékoli fotografie). Pan Neubert se vyjádřil, že digitalizace není vnímána jako prostředek pro zajišťování bezpečných kopií. Chápe složitost problému a uvědomuje si, že nikdo nechce říct jasné a definitivní stanovisko. FADGI je jistě dobrý projekt, ale je považován za experiment, a proto jeho výsledky nelze vnímat jako všeobecně přijímané. Např. mapové oddělení vytváří tištěné kopie, které opatřuje analogovým razítkem. Tedy z pohledu autenticity kopií je tato otázka značně otevřená různým přístupům. Na otázku k přístupu vedení LOC k nastavování strategií v oblastech péče o uchovávané sbírky bylo uvedeno, že vedení nemá samo o sobě příliš rádo mikrofilmovou technologii (nepovažuje ji za moderní). V LOC se mikrofilmují především noviny, protože je to jednoduché a asi 3x levnější než digitalizace. Audio nahrávky se naopak uchovávají pouze v digitální podobě (na CD nebo DVD). Pan Neubert připouští, že

digitální strategie se těžko rodí v tak obrovské instituci jakou je LOC, dokonce i koordinace digitálních projektů je zde náročná práce. Přesto prý nedochází k duplicitní digitalizaci.

Dále jsme s panem Neubertem navštívili digitalizační středisko, kde se skenují všeobecné sbírky v rámci projektu Internet Archives. Jedná se o rutinní skenovací pracoviště. Zajímavé bylo zjištění, že u snímků vytvořených pomocí kamery, ručně přepočítávají optické rozlišení podle stanoveného ohniska. Výstupy ukládají do formátu JPEG2000. Bylo nám sděleno, že přibližně 30% naskenovaných dokumentů lze paralelně najít na Internetu mezi digitalizovanými dokumenty spol. Google, která skenuje masově bez kvalitativních měřítek. Žádnou spolupráci se spol. Google nerozvíjí a LOC není aktivitám spol. Google příliš nakloněná.

Závěr naší návštěvy v LOC byl vyhrazen finální diskuzi s koordinátorem výjezdu panem Pugliou nad otázkami přístupu LOC k mikrofilmování a digitalizaci a jejich vzájemnému vztahu. Pan Puglia přiznal, že zatím neexistuje žádné oficiální doporučení v LOC k digitalizaci, která by byla určená výhradně pro účely uchování. IT odborníci toto doporučení podporují, zatímco tradiční knihovníci digitalizaci v tomto kontextu nevnímají. Množství dostupných finančních prostředků samozřejmě ovlivňuje rozsah a kvalitu digitálních výstupů, která pak dále ovlivňuje jejich využití. Vedení knihovny přitom stále častěji upřednostňuje digitalizaci před mikrofilmováním. Co se týká datových formátů, několikrát byl zmíněn JPEG2000 a jeho použití v LOC. Dále jsou používány TIFF, JPEG, PNG, PDF (PDF/A). Žádný z uvedených formátů však není preferován na úkor ostatních, třebaže nejvíce se v LOC používá formát TIFF.

Shrnutí výsledků zahraniční cesty

Zahraniční cesta poskytla projektovému týmu zajímavý pohled na řešení problémů souvisejících s mikrofilmováním a digitalizací ve dvou předních paměťových institucí v USA i ve světě. Minulé i současné možnosti NARA i LOC v oblastech finančních, technologických a personálních zdrojů umožnili přistupovat k otázce vytváření zajišťovacích (bezpečnostních) kopií značně velkoryse a nekoordinovaně. Bude zajímavé sledovat, jak se tento přístup bude vyvíjet a měnit s ohledem na plánované úspory státní správy a další vývoj v oblasti informačních technologií.