Dlouhodobá archivace digitálních dat – od teoretických úvah k praktické realizaci?

Digital Preservation – from theory to practice?

Mgr. Jan Hutař, Ph.D. /Archives New Zealand10 Mulgrave Street, Thorndon, Wellington 6011, New Zealand ; Mgr. Marek Melichar / Ústav výpočetní techniky, Univerzita Karlova v Praze (Computer Science Centre, Charles University in Prague) , Ovocný trh 3/5, 116 43 Praha 1, Česká republika

Resumé:
Článek stručně informuje o tématu dlouhodobé archivace digitálních informací (digital preservation). Po úvodu do problematiky a vysvětlení základních termínů jsou krátce popsány koncepty referenčního rámce OAIS (ČSN ISO 14721), základní přístupy k dlouhodobé archivaci. Jsou zmíněny cíle a některé závěry projektu LTP Pilot, ve kterém byl testován systém Archivematica.

Klíčová slova: dlouhodobá archivace, OAIS, migrace, digitální data, paměťové instituce

Summary:
This article brings introductory information about the long-term digital preservation. The text explains basic terminology and concepts introduced in the OAIS reference model in the functional and information model. Then it describes the pragmatic approach to the long-term archiving. Finally it mentions some of the existing technological solutions as well as findings of the LTP Pilot project in which Archivematica system was evaluated.

Keywords: digital preservation, OAIS, migration, digital data, culture heritage institutions

Tento příspěvek byl vytvořen v rámci řešení projektu 516R1/2014 „Pilotní projekt pro low-barrier přístup k ochraně digitálního obsahu (LTP-pilot)“ financovaného Fondem rozvoje CESNET

1 Co je dlouhodobá archivace digitálních dat

Naše schopnost porozumět informačnímu obsahu digitálních objektů závisí nejen na našich znalostech, ale z velké části na technologiích, které k nalezení a použití digitálních dat potřebujeme. Zastarávání programů (software) a technického vybavení (hardware), riziko ztráty kontextu nebo ztráty možnosti objekty vyhledat a identifikovat jsou spolu s celkovou zranitelností digitálních objektů těmi problémy, které je nutno řešit. V posledních letech narůstá objem digitálních dat vytvářených a sdílených ve všech oblastech našeho života. Některá (zdaleka ne všechna, viz např. van der Werf, T., & van der Werf, B.[1]) vznikající data je třeba uchovat pro budoucí použití. Ve

[1] WERF, Titia van der a Bram van der WERF. *The paradox of selection in the digital age*. In: *IFLA WLIC 2014, 16–22 August 2014, Lyon, France*.
Dostupné také z: *http://library.ifla.org/1042/1/138-vanderwerf-en.pdf*.

chvíli, kdy jde o velké objemy dat, jež jsou například výsledkem několikaletého úsilí vědeckého výzkumu či skenování fyzických předloh v paměťových institucích, dále v případech, kdy dokumenty už nemají fyzické předlohy nebo jde o dokumenty kritické pro bezpečnost či zdraví lidí, dokumenty vznikající ve státní správě atp., je naše schopnost trvale uchovávat obsah v digitální podobě velmi důležitá.

Po sítích sdílíme obrovské objemy dat[2], dokážeme obrovské objemy dat ukládat, ale podle odhadů jen polovina informací, které by měly být zabezpečeny, také nějaké zabezpečení skutečně má[3]. Fyzické uchování dat v technologicky použitelné podobě nemusí stačit. Dlouhodobá archivace informačních obsahů v digitální podobě je systematickou a praktickou činností směřující k jedinému cíli: uchovat intelektuální obsah kódovaný v digitálních datech v takové podobě, aby tyto informace byly použitelné a dávaly smysl a i budoucím generacím uživatelů.

Nejde tedy o nějakou konceptuální nebo filosofickou disciplínu, ale o zcela prakticky orientovanou každodenní činnost mnoha lidí v paměťových institucích i firmách. Jejím cílem je: „*uchování obsahu pro budoucí použití a zpřístupnění*“[4]. Dlouhodobá archivace má další neméně důležité a možná často nepřímé cíle, jako je budování sbírek, podpora smyslu instituce, vyhovění zákonným požadavkům, ochrana investic vložených do pořízení digitálních dat, podpora důvěryhodnosti instituce apod.[5] – podobně jako v papírovém světě.

Kolem dlouhodobé archivace digitálních informací koluje řada mýtů a nedorozumění, které souvisejí s komplexní povahou problematiky, nejednoznačností terminologie a některých konceptuálních standardů. Také stále málo institucí dlouhodobou archivaci skutečně prakticky provádí. Vytváření obrazu dlouhodobé archivace digitálních informací jako teoretické a konceptuální disciplíny nebo jako činnosti vyžadující vysoce odborné a technické znalosti ztěžuje mnoha institucím rozhodnutí, zda a jak se ochranou digitálních dat zabývat. Proto vznikají iniciativy podporující digitální archivaci osobních dat (tzv. *personal digital archiving*), které vydávají různé návody a metodické publikace[6] nebo pořádají konference[7]. Z podobných důvodů existují také iniciativy podporující snížení bariér dlouhodobé archivace v institucích, např. projekt Preserving digital Objects with Restricted Resources (POWRR – Ochrana digitálních objektů s omezenými zdroji[8]) nebo u nás Projekt pro low-barrier přístup k ochraně digitálního obsahu (LTP-pilot)[9], který je podrobněji popsán dále v textu.

---

2 CISCO. The Zettabyte Era – Trends and Analysis, White Paper. Cisco.com [online]. May 2015. [cit. 2015-07-28]. Dostupné z: *www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/VNI_Hyperconnectivity_WP.html.*

3 GANTZ, John a David REINSEL. *The 2011 Digital Universe Study: extracting Value from Chaos* [online]. IDC, 2011 [cit. 2015-07-28]. s. 1. Dostupné z: *http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf.*

4 LIBRARY OF CONGRESS. About – Digital Preservation (Library of Congress) [online]. [s.a.] [cit. 2015-10-03]. Dostupné z: *http://www.digitalpreservation.gov/about/.*

5 BROWN, Adrian. *Practical digital preservation: a how-to guide for organizations of any size*. 1st ed. London: Facet, c2013, xvi, 336 s. ISBN 978-1-85604-755-5.

6 Např. návody na stránkách Kongresové knihovny zde: *http://www.digitalpreservation.gov/personalarchiving/.*

7 Např. letošní konference Personal Digital Archiving se konala v New Yorku, viz *http://personaldigitalarchiving.com/.*

8 *http://digitalpowrr.niu.edu/*

9 *http://ltp-portal.cz* a *http://www.muni.cz/ics/research/projects/28563*

Dlouhodobá archivace není jen zálohování a udržování „více kopií na více místech“ nebo něco, čemu bychom se mohli věnovat až někdy později (Corrado, E. M., and H. L. Moulaison[10]). Také to není aktivita, kterou bychom mohli nechat jen na bedrech národních institucí, ani aktivita na kterou jsou třeba obrovské finanční prostředky, nebo která nutně vyžaduje velmi expertní technické znalosti. A. Brown v publikaci Practical digital preservation: a how-to guide for organizations of any size (Praktická dlouhodobá archivace: návod pro libovolně velkou organizaci[11]) ukazuje, jak organizace jakékoli velikosti může začít i ve skromných podmínkách pracovat na tom, aby jí spravovaná data byla dlouhodobě použitelná.

2 Pragmatický přístup k dlouhodobé archivaci

Dlouhodobé uchování informací v digitální podobě (nebo trvalé uchování) vyžaduje nejen ochranu „jedniček a nul“, tedy bitového streamu (sekvence bitů). Ochrana bitového streamu neřeší problémy se zastaráváním softwaru, hardwaru a formátů vlastních digitálních objektů, ani problémy s autenticitou a použitelností intelektuálního obsahu. Pasivní uchování bitů je pouze *předpokladem*, případně prvním krokem *ochrany logické*. Logická dlouhodobá ochrana digitálních dat je aktivní, systematická a plánovaná. Spočívá v činnostech prováděných během životního cyklu digitálního objektu tak, aby byla zajištěna trvalá použitelnost informačního obsahu. Použitelnost je obecný pojem pro vyhledatelnost, zobrazitelnost, pochopitelnost a autentičnost obsahu. K zajištění trvalé použitelnosti musí být dokumenty v archivu stále „živé“, musí reflektovat změny v globálním technickém prostředí a reagovat na změny, které vyžaduje správa dokumentů[12]. Vedle ochraňovaných objektů musí být uchovávána také odpovídající metadata (o vlastnostech, kontextu, původu, krocích během archivace aj.), která jsou v průběhu životního cyklu neustále doplňována. Každá událost či změna archivního objektu má být zaznamenána v metadatech.

V praktické rovině je krátkodobým cílem předání dat další generaci s dostatečnými doprovodnými metadaty a informacemi o vlastních datech, a také o změnách a kontrolách provedených nad uloženými daty během dosavadní archivace. Trvalé uchovávání vyžaduje pragmatická rozhodnutí o tom, jaké typy informací a událostí zaznamenávat, v jakém formátu, jaké technické informace z objektů extrahovat, jaké kontextové a vysvětlující informace přidávat apod. Pragmatický přístup znamená přistupovat k dlouhodobé archivaci s postupně rostoucími ambicemi, a nikoli očekávat trvalé a dokonalé řešení hned na začátku, nebo případně nedělat nic. Zmíněný projekt POWRR popsal minimální požadavky pragmatického přístupu k dlouhodobé archivaci takto[13]:

---

10 CORRADO, E. M. a H. L. MOULAISON. *Digital preservation for libraries, archives, and museums*. Lanham, MA: Rowman & Littlefield, 2014. 270 s. ISBN 0810887126.

11 BROWN, Adrian. *Practical digital preservation: a how-to guide for organizations of any size*. 1st ed. London: Facet, c2013, xvi, 336 s. ISBN 978-1-85604-755-5.

12 FOJTŮ, Andrea, Jan HUTAŘ a Marek MELICHAR. Dlouhodobá ochrana digitálních dokumentů a projekt NDK. In: *Knihovny současnosti 2011: Sborník z 19. konference, konané ve dnech 13.–15. září 2011 v Českých Budějovicích*. Ostrava: Sdružení knihoven ČR, 2011, s. 73–79. ISBN 978-80-86249-62-9.
Dostupné také z: *http://www.sdruk.cz/data/xinha/sdruk/ks2011/sbornik_2011.pdf*.

13 MINER, M. From Theory to Action: A Pragmatic Approach to Digital Preservation Strategies and Tools. In: *SAA Research Forum, Joint Annual Meeting of CoSA, NAGARA, and SAA, Washington, DC. Aug. 10–16, 2014*.
Dostupné také z: *http://powrr-wiki.lib.niu.edu/images/6/69/POWRR_outcomes.pdf*.

– vytvoř inventář obsahu a analyzuj obsah s využitím nástrojů jako je NDSA's Levels of Preservation[14] (viz níže),
– využij jednoduché nástroje pro získání metadat, vytvoření balíčků a zajištění dat na úložišti,
– sleduj a více využívej robustnějších nástrojů,
– uvědom si, že plánování a zajištění financí je stejně důležité jako technologie,
– sleduj, co se děje v komunitě, využívej zdroje vytvořené komunitou.

Přístup POWRR tedy mj. zdůrazňuje skutečnost, že instituce nejsou ve své snaze uchovávat data v digitální podobě nikdy samy – existuje rozsáhlá informační infrastruktura a komunitou sdílené nástroje, které jsou prakticky využitelné i v menších projektech s malými rozpočty.

Prokázání kvality dlouhodobé archivace je možné s využitím řady nástrojů pro audit, vnitřní audit (self audit) a certifikaci. Existují i řešení snadno dostupná a použitelná pro menší instituce. V některých institucích je například aplikace standardů, jako je ISO 16363[15] nebo ISO 27000[16] apod. nereálná. Existují ale nástroje, které se dají použít při plánování a prokazování kvality v jakékoliv instituci (DSA, Nestor seal, Platter, DRAMBORA, NDSA's Levels of Preservation[17]), a přitom nevyžadují velké náklady. I jejich aplikace bezpochyby zvyšuje důvěryhodnost projektu dlouhodobé archivace.

Pragmatický a praktický přístup k dlouhodobé archivaci řeší obtížné požadavky na dlouhodobou archivaci pragmatickými rozhodnutími v konkrétních projektech s ohledem na dostupné zdroje, finance, a technologie. Samozřejmě se snaží stále zajistit, aby intelektuální obsah uložených objektů zůstal nezměněný a autentický a byl zároveň použitelný v budoucím neznámém technologickém prostředí uživateli, o jejichž znalostech nic nevíme, byl použitelný i bez původce dat atd. Nicméně každodenní provoz archivu a reálná data mohou vyžadovat pragmatická nebo dočasná řešení. Stejně tak instituce musí řešit výběr dokumentů k dlouhodobé archivaci – bohužel paměťové instituce nikdy nebudou disponovat takovou kapacitou, aby mohly ukládat vše, co by z digitálního světa ukládat chtěly[18].

Příkladem jednoduchého návodu pro posouzení úrovně zajištění dlouhodobé archivace v instituci může být model NDSA vyvinutý stejnojmennou americkou organizací (NDSA – National Digital Stewardship Alliance[19]). Model má identifikovat problémy

---

14 LIBRARY OF CONGRESS. NDSA Levels of Preservation – NDSA – Digital Preservation (Library of Congress). [online]. [s.a.] [cit. 2015-10-03].
Dostupné z: *http://www.digitalpreservation.gov/ndsa/activities/levels.html.*

15 ISO 16363:2012. *Space data and information transfer systems -- Audit and certification of trustworthy digital repositories*. Geneva: International Organization for Standardization, 2012. 70 s.

16 ISO 27000:2014. *Information technology -- Security techniques -- Information security management systems -- Overview and vocabulary*. Geneva: International Organization for Standardization, 2014. 31 s.

17 Data Seal of Approval – *http://datasealofapproval.org/en/,* případně český překlad na *http://dsa.cuni.cz/*; Nestor Seal – *http://www.langzeitarchivierung.de/Subsites/nestor/EN/nestor-Siegel/siegel_node.html*, Platter – dostupný česky na *http://www.ndk.cz/platter-cz/Platter.pdf*; DRAMBORA – *http://www.repositoryaudit.eu/*; NDSA Levels of Preservation – *http://www.digitalpreservation.gov/ndsa/activities/levels.html*

18 WERF, Titia van der a Bram van der WERF. *The paradox of selection in the digital age*. In: *IFLA WLIC 2014, 16–22 August 2014, Lyon, France*. Dostupné také z: *http://library.ifla.org/1042/1/138-vanderwerf-en.pdf.*

19 LIBRARY OF CONGRESS. NDSA Levels of Preservation – NDSA – Digital Preservation (Library of Congress). [online]. [s.a.] [cit. 2015-10-03].
Dostupné z: *http://www.digitalpreservation.gov/ndsa/activities/levels.html.*

a tedy pomoci určit priority dalšího rozvoje. Stručně a přehledně popisuje čtyři úrovně dlouhodobé ochrany digitálních informací v pěti oblastech[20]. Všechny prvky první úrovně jsou předpokladem pro budování systémů a procesů v úrovni navazující. Plán dlouhodobé ochrany digitálních dat může být rozepsán podrobněji pro každou oblast, podle možností a cílů instituce. Úrovně mohou být aplikovány na konkrétní sbírky nebo na celé systémy ukládající více digitálních sbírek.

Tab. 1 Úrovně dlouhodobé ochrany digitálních dokumentů podle NDSA, OWENS, 2012

		Úrovně				
		Úroveň nultá (ukaž data)	**Úroveň první (ochraň data)**	**Úroveň druhá (poznej data)**	**Úroveň třetí (monitoruj data)**	**Úroveň čtvrtá (oprav data)**
Oblasti	**Datové úložiště a jeho geografické umístění**	Data jsou někde uložena na nějakých médiích.	Existují dvě úplné kopie dat, které nejsou umístěné na stejném místě. Data z heterogenních nosičů (optické disky, přenosné harddisky, apod.) je nutné přenést do datového úložiště.	Ukládají se tři úplné kopie dat. Alespoň jedna kopie v jiné lokalitě. Jsou dokumentovány systémy úložiště dat a úložná média včetně informací o všem, co je třeba k jejich použití.	Ukládají se nejméně tři úplné kopie dat. Alespoň dvě kopie dat se nacházejí v lokalitách, které nesdílejí shodný druh ohrožení (např. přírodních katastrof, ale znamená to i různý hardware a souborový systém). Existuje proces sledování zastarávání úložných systémů a médií.	Ukládají se nejméně tři úplné kopie dat. Tři kopie dat se nacházejí každá v jiné lokalitě, žádné dvě z těchto lokalit nesdílejí shodný druh ohrožení. Je vypracován podrobný plán, který zajistí, že soubory i metadata jsou uloženy na dostupných médiích nebo systémech.
	Integrita dat a neměnnost souborů	Není známa / nelze zkontrolovat.	Kontrola integrity souborů probíhá při převzetí dat, pokud byla data dodána s kontrolním součtem. Pokud nejsou kontrolní součty součástí dodávky dat, jsou při převzetí dat vytvářeny.	Kontrola integrity u všech přebíraných dat. Originální média se blokují proti zápisu. Vysoce rizikový obsah prochází antivirovou kontrolou.	Integrita dat je kontrolována v pravidelných intervalech. Udržují se záznamy (logy) o stavu integrity dat, na požádání lze dodat audit těchto informací. Lze detekovat poškozená data. Všechen obsah prochází antivirovou kontrolou.	Kontroluje se integrita všech dat v návaznosti na konkrétní události nebo aktivity. Zajistit, že žádná osoba nemá právo zápisu ke všem kopiím dat.
	Informační bezpečnost	Není známa.	Ví se, kdo má práva čtení, přesouvání a mazání souborů. Omezit tato přístupová oprávnění k jednotlivým souborům.	Přístupová oprávnění k obsahu jsou zdokumentována.	Jsou uchovávány záznamy (logy) toho, kdo prováděl jaké akce s jakými soubory, včetně mazání a akcí digitální ochrany.	Jsou prováděny audity těchto záznamů.

[20] OWENS, T. NDSA Levels of Digital Preservation: Release Candidate One. In: The Signal: Digital Preservation [online]. Library of Congress, 2012 [cit. 2015-10-03]. Dostupné z: *http://blogs.loc.gov/digitalpreservation/2012/11/ndsa-levels-of-digital-preservation-release-candidate-one/.*

Metadata	Nějaká metadata lze odvodit z názvů souborů, jejich atributů a adresářové struktury.	Existuje přehled obsahu repozitáře a jeho konkrétního umístění na úložišti. Je zajištěno zálohování tohoto seznamu a jeho záloha v jiné lokalitě. Je zaveden lokální jednoznačný identifikátor.	Jsou ukládány administrativní metadata. Ukládají se metadata o transformacích a záznamy událostí.	Ukládají se standardní technická a popisná metadata.	Ukládají se standardní ochranná metadata. Je zaveden globálně jednoznačný identifikátor.
Formáty souborů	Dají se odvodit z přípony souborů.	Pokud je to možné ovlivnit, podporujte používání malé skupiny dobře známých a otevřených formátů souborů a kodeků.	Existuje seznam formátů, které jsou používány.	Monitorují se hrozby zastarávání formátů.	Provádí se formátové migrace, emulace a podobné aktivity podle potřeby.
Práva	Je akceptován fakt, že je nutné mít ujasněny zodpovědnosti za ochranu a právní vztahy k digitálním objektům, které mají být předmětem ochrany.	Odpovědnost za ochranu je vyjasněná - máme právo daná data trvale uchovávat.	Je známo, co je možné s předmětem ochrany dělat z hlediska použití a zpřístupnění.	Existuje oprávnění provádět akce digitální ochrany (např. migraci do nových formátů).	Existuje oprávnění vytvářet a zpřístupňovat odvozené dokumenty pro definovanou komunitu. Právo přenést zodpovědnost za ochranu a práva k obsahu na někoho jiného.

3 Referenční rámec OAIS

Základem terminologie a praktických návrhů dlouhodobé archivace je referenční rámec OAIS (ČSN ISO 14721[21]). Koncepty obsažené v OAIS jsou komplexnější, než přístup modelu NDSA, i přesto je OAIS široce akceptován. OAIS má za sebou již 20 let existence – v polovině 90. let 20. století archivní a knihovnické komunitě chyběl obecný rámec, který by popsal společná východiska a jazyk pro další vývoj archivačních systémů pro digitální data. Práce CCSDS (*Consultative Committee for Space Data Systems),* která se ujala vytvoření standardu, vyústila v květnu 1999 ve vydání referenčního modelu OAIS, který byl v roce 2003 publikován jako mezinárodní norma ISO 14721:2003[22]. Poslední, doplněná verze normy, z roku 2012, existuje i v českém překladu[23]. Referenční

[21] ČSN ISO 14721. *Systémy pro přenos dat a informací z kosmického prostoru - Otevřený archivační informační systém - Referenční model*. Praha: Úřad pro technickou normalizaci, metrologii a státní zkušebnictví, 2014.

[22] ISO 14721:2003. *Space data and information transfer systems -- Open archival information system -- Reference model*. Geneva: International Organization for Standardization, 2003.

[23] ISO 14721:2012. *Space data and information transfer systems – Open archival information system (OAIS) – Reference model*. Geneva: International Organization for Standardization, 2012. 126 s. a ČSN ISO 14721. *Systémy pro přenos dat a informací z kosmického prostoru – Otevřený archivační informační systém – Referenční model*. Praha: Úřad pro technickou normalizaci, metrologii a státní zkušebnictví, 2014.

model OAIS se ukázal jako velmi životaschopný a je dnes široce implementován a využíván v paměťových institucích i v komerčních řešeních[24]. Je ideální svou obecností a tím i možností implementace. Velmi podstatné je, že referenční rámec OAIS obsahuje a definuje:

- Terminologický slovník pro oblast dlouhodobé ochrany a archivů/repozitářů, který je srozumitelný široké odborné veřejnosti.
- Informační model, který především popisuje podrobně strukturu a obsah archivního informačního balíčku AIP (Archival Information Package), tj. říká, jaká metadata mají být spolu s uchovávaným obsahem ukládána. AIP obsahuje kompletní sadu nazvanou ***Popisné ochranné informace*** (***Preservation Description Information – PDI***[25]) pro obsahové informace.
- Funkční model digitálního archivu. Klíčové funkce dlouhodobého archivu budovaného v souladu s koncepty OAIS jsou popsány v šesti funkčních celcích – **Příjem (*Ingest*), Archivní uložení (*Archival Storage*), Správa dat (*Data Management*), Správa (*Administration*), Zpřístupnění (*Access*)** a **Plánování uchovávání (*Preservation Planning*).** Tedy popisuje klíčové procesy probíhající v digitálním archivu a jeho funkční komponenty.

Současný stav standardu OAIS a jeho minulý i současný vliv na komunitu dlouhodobé archivace je popsán ve druhé edici práce B. Lavoie *The Open Archival Information System (OAIS) Reference Model: Introductory Guide*[26]. Lavoie mj. popisuje vývoj dalších standardů používaných v oblasti dlouhodobé archivace, které z OAIS vycházejí nebo jsou modelem OAIS inspirované (především metadatové – de facto – standardy jako METS[27], PREMIS[28] nebo další standardy ISO jako ISO 16363 a ISO 16919[29], sloužící pro prokázání kvality v oblasti dlouhodobé archivace).

4 Prakticky použitelné nástroje

V posledních 15 letech vzniklo v institucích a projektech zabývajících se dlouhodobou archivací z výzkumného nebo praktického hlediska obrovské množství nástrojů, metodik a příkladů dobré praxe. Řada z těchto řešení je dostupná volně a dlouhodobou archivaci si bez nich mnoho institucí nedovede představit. Instituce běžně používají volně dostupné nástroje operačních systémů nebo nástroje jako jsou Imagemagick,

[24] FERLE, Christoph H. Marktstudie digitale Langzeitarchivierung: im Spannungsfeld zwischen Digital Preservation und Enterprise Information Archiving. Stuttgart: Fraunhofer, 2012. Dostupné také z: *http://www.swm.iao.fraunhofer.de/content/dam/swm/de/documents/Marktstudie_Digitale_Langzeitarchivierung_web.pdf*.

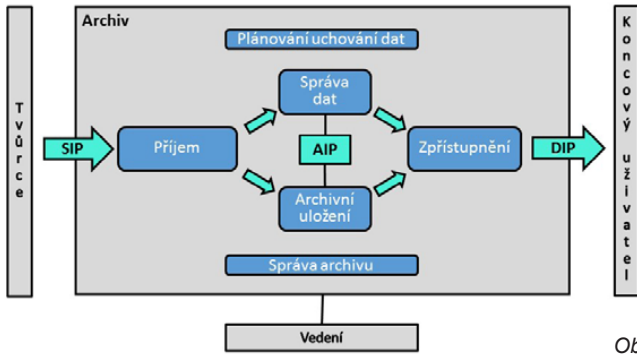
[25] Terminologie z ISO 14721:2012.

[26] LAVOIE, Brian. *The Open Archival Information System (OAIS) Reference Model: Introductory Guide (2nd Edition)*. Digital Preservation Coalition, 2014. 33 s. Dostupné z: *http://dx.doi.org/10.7207/TWR14-02*.

[27] METS – Metadata Encoding and Transmission System, viz *http://www.loc.gov/standards/mets/*.

[28] PREMIS – Preservation Metadata: Implementation Strategies, viz *http://www.loc.gov/standards/premis/*.

[29] ISO 16919:2014. *Space data and information transfer systems -- Requirements for bodies providing audit and certification of candidate trustworthy digital repositories*. Geneva: International Organization for Standardization, 2014. 22 s.



Obr. 1 Referenční rámec OAIS[30]

Ghostscript, Open office[31] a další pro migrace formátů, OCR, generování hashů apod. Specifické nástroje jako jsou DROID, FIDO, SIEGFRIED, JHOVE a JHOVE2, FITS, Jpylyzer, mediaInfo, ffprobe, Exiftool, New Zealand Metadata Extraktor, Apache Tika, PDFbox, VeraPDF[32] jsou klíčové pro provoz mnoha dlouhodobých archivů při validaci formátů, extrakci technických metadat, identifikaci formátů. Pro podporu plánování dlouhodobého uchovávání existuje řada příkladů strategií a politik, nástrojů jako jsou test beds apod.

Kompletní řešení a systémy pro dlouhodobou archivaci (LTP – Long Term Preservation) vznikají od okamžiku formulace první verze modelu OAIS. První generace systémů pro logickou ochranu digitálních dat se objevila na začátku nového tisíciletí. Mezi prvními implementovala takový systém nizozemská národní knihovna s produktem IBM DIAS[33]; DAITSS byl vyvíjen v souvislosti s repozitářem *Florida Digital Archive*[34] od roku 2005; XENA[35] byl projekt a systém Australského národního archivu. LTP systémy druhé generace vznikající koncem prvního desetiletí 21. století jsou již pokročilejší (např. Rosetta, Preservica, Archivematica, RODA[36]); poučily se z pionýrských začátků a využívají také výsledky výzkumných projektů financovaných ze zdrojů EU nebo z veřejných zdrojů

30 CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEMS. *Reference Model for an Open Archival Information System (OAIS): CCSDS 650.0-B-1* [online]. Washington (DC): Consultative Committee for Space Data Systems, January 2002 [cit. 2015-07-28]. s. 4-1. Dostupné z: *http://public.ccsds.org/publications/archive/650x0b1.PDF.*

31 Imagemagick – *http://www.imagemagick.org/script/index.php* ; Ghostscript – *http://www.ghostscript.com/* ; Open office – *https://www.openoffice.org/.*

32 DROID – *http://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/* ; FIDO – *http://openpreservation.org/technology/products/fido/* ; SIEGFRIED – *http://www.itforarchivists.com/siegfried* ; JHOVE – *http://jhove.sourceforge.net/* ; JHOVE2 – *https://bitbucket.org/jhove2/main/wiki/Home* ; FITS – *http://projects.iq.harvard.edu/fits/fits-xml* ; Jpylyzer – *http://jpylyzer.openpreservation.org/* ; MediaInfo – *https://mediaarea.net/cs/MediaInfo* ; ffprobe – *https://ffmpeg.org/ffprobe.html* ; ExifTool – *http://www.sno.phy.queensu.ca/~phil/exiftool/* ; New Zealand Metadata Extraktor – *http://meta-extractor.sourceforge.net/* ; Apache Tika – *https://tika.apache.org/* ; Apache PDFbox – *https://pdfbox.apache.org/* ; VeraPDF – *http://openpreservation.org/about/projects/verapdf/*

33 DIAS – *http://www-935.ibm.com/services/ch/gts/pdf/br-storage-lza-en-01-04-08.pdf*

34 DAITSS – *https://daitss.fcla.edu/*

35 XENA – *http://xena.sourceforge.net/*

36 Rosetta – *http://www.exlibrisgroup.com/category/RosettaOverview* ; Preservica – *http://preservica.com/* ; Archivematica – *https://www.archivematica.org/en/* ; RODA – *http://www.roda-community.org/*

v USA, Velké Británii a jinde (SCAPE, Planets, NDIIPP[37] atd.). Po roce 2009 se začaly objevovat také open source systémy s ambicí naplnit kompletně požadavky konceptů OAIS.

Řada institucí buduje svoje LTP systémy na míru z dostupných open source komponent, které spojují do funkčních celků. K takovýmto systémům je obtížné získat dokumentaci a nejsou obvykle dostupné jako produkty nebo kompletní řešení (např. národní knihovny Francie, Nizozemí, Švýcarska, Dánska aj.). V oblasti firemní archivace existuje snaha rozšiřovat funkcionality klasických systémů ECM (Electronic Content Management) o funkce požadované v oblasti dlouhodobé archivace (viz Korb, J., & Strodl[38] nebo Ferle[39]).

5 Projekt LTP Pilot

V posledních dvou letech se stal populárním LTP systém Archivematica. Tento systém byl také předmětem testování v projektu 516R1/2014 „Pilotní projekt pro low-barrier přístup k ochraně digitálního obsahu (LTP-pilot)“ financovaném Fondem rozvoje CESNET, jehož se účastnili vedle řešitele (ÚVT Masarykovy Univerzity) také další instituce (Moravská zemská knihovna, Jihočeská univerzita). Také některé další instituce v České republice v posledních dvou letech experimentovaly nebo stále experimentují s tímto systémem (například Knihovna Akademie věd). Systém Archivematica používá v kontextu svého právě vyvíjeného řešení také Národní archiv. Jde o perspektivní systém, jehož vývoj je řízen firmou Artefactual[40] ve spolupráci s UNESCO a řadou severoamerických univerzit a dalších paměťových institucí v Evropě.

Záměrem projektu LTP Pilot bylo testování funkcionality systému Archivematica s využitím infrastruktury digitálních úložišť CESNET, získání kurátorských zkušeností se zpracováním dat ve worfklow systému a s jeho konfigurací a vkládáním testovacích sbírek dat z různých zdrojů, posouzení systému z hlediska vyhovění konceptům OAIS, a také posouzení robustnosti a škálovatelnosti systému.

Archivematica splňuje především část funkcí OAIS z oblasti příjmu dat (ingest) a archivního úložiště (archvival storage). Testování ukázalo, že systém sice zatím není z pohledu funkcionality a výkonu příliš robustní a dostupná dokumentace k němu není dokonalá, dokáže však ve workflow spojujícím jednotlivé mikroslužby zpracovat bez problémů různé typy dat, ať již vkládaných bez metadat, nebo s metadaty, případně z již hotových balíčků SIP dokáže vytvořit balíčky AIP, které obsahují všechny informace požadované informačním modelem OAIS. Archivematica používá v této oblasti obvyklé standardy (METS, PREMIS, Dublin Core) a balíčky AIP „balí“ do kontejneru BagIt[41]. Při testování zpracování dat z různých sbírek byly využity existující mechanismy transferu dat, a také byly vytvořeny skripty pro přípravu dat do standardního formátu pro transfer. Byly jednak

---

37 SCAPE – *http://www.scape-project.eu/*; Planets – *http://www.planets-project.eu/*; NDIIPP (National Digital Information Infrastructure and Preservation Program) – *http://www.digitalpreservation.gov/*

38 KORB, J., & S. STRODL. Digital preservation for enterprise content: a gap-analysis between ECM and OAIS. In *7th International Conference on Preservation of Digital Objects, Vienna* (pp. 221–228). Též česky na: *https://duha.mzk.cz/clanky/dlouhodoba-ochrana-podnikovych-dokumentu-analyza-rozdilu-mezi-ecm-oais.*

39 FERLE, Christoph H. Marktstudie digitale Langzeitarchivierung: im Spannungsfeld zwischen Digital Preservation und Enterprise Information Archiving. Stuttgart: Fraunhofer, 2012. Dostupné také z: *http://www.swm.iao.fraunhofer.de/content/dam/swm/de/documents/Marktstudie_Digitale_Langzeitarchivierung_web.pdf.*

40 *https://www.artefactual.com/*

41 *https://en.wikipedia.org/wiki/BagIt*

využity již zapojené nástroje třetích stran pro zpracování formátů (již zmíněné nástroje FIDO, FITS, JHOVE), jednak byly testovány mechanismy zapojení dalších nástrojů.

Archivematica není repozitář v tom smyslu, v jakém repozitář obvykle chápeme. Není to tedy systém, který by umožňoval v současné době s vytvořenými a uloženými balíčky AIP nějak dále pracovat nebo je přímo zpřístupňovat koncovým uživatelům. Tyto funkce je třeba implementovat vně systému Archivematica (zpřístupnění) nebo lze využívat API systému v kombinaci s jinými systémy na správu dat (aktualizace AIP a jejich metadat). Instituce často hledají systémy, které jim umožní udržovat pořádek v archivovaných datech a pracovat s nimi. Potřebují nejen systémy, které jim umožní naplňovat základní požadavky OAIS, ale také systémy pro efektivní správu archivovaného obsahu. V této oblasti má Archivematica stále velké mezery, které je nutno vyplnit jiným systémem. Kompletní řešení pro dlouhodobou archivaci a zpřístupnění budují současní uživatelé systému Archivematica s využitím dalších systémů jako je DSPACE, DURASPACE, Islandora, iRods[42], za využití technologie Hierarchical Storage Management apod.

Závěr

Dlouhodobá archivace digitálních dat je oblastí, která se neustále vyvíjí. Některé skutečnosti zůstávají neměnné, ale nástroje, systémy a technologie se mění relativně rychle. Vývoj dlouhodobé archivace je vlastně neustálou změnou. Samotná problematika se zrodila z akutní potřeby vypořádat se se specifickou podstatou digitálních dokumentů, která je odlišná od fyzických dokumentů, a s problémy, které přináší oproti fyzickým dokumentům. Vznikly standardy (OAIS), první jednoduché nástroje (JHOVE) a později i komplexní LTP systémy. Současně byly řešeny projekty, ať evropské nebo jiné, které posouvaly vývoj dále. Digitální archivace se jako disciplína ustálila na konci minulého tisíciletí, instituce se jí začaly věnovat, začaly se provádět audity existujících digitálních repozitářů, certifikace. Digitální archivace se mnohde stala rutinou. Mnohé instituce ale stále tápají.

Smyslem tohoto textu je upozornit na praktické a pragmatické možnosti v oblasti dlouhodobé archivace. Nemá smysl čekat na dokonalá řešení nebo na velké finanční prostředky. Naskýtá se otázka, zda dokonalá řešení budou vůbec kdy existovat; dosáhneme-li jednou dokonalých řešení, může být ale část našich dat již dávno ztracena.

Akademický diskurs o dlouhodobé archivaci, kde se debatuje o významu terminologie, smyslu konceptů OAIS a vazbách na jiné koncepty, často bez reálné zkušenosti z každodenní správy a z péče o uchovávání a zpřístupňování konkrétních digitálních dat, nijak paměťovým institucím nepomáhá. Pomoci jim mohou jednoduché nástroje a příklady architektur řešení využívající volně dostupné komponenty[43]. Dlouhodobá archivace digitálních informací je nutně komunitní aktivitou, žádná instituce nebo jednotlivec se neobejde bez informací a nástrojů poskytovaných různými komunitami technických expertů, ani bez podpory systémových správců, správců hardwaru a úložišť a programátorů, ani bez pochopení managementu institucí nebo politické podpory projektů. Klíčové je sdílení informací a poznatků „napříč“ institucemi i státy.

Dlouhodobá archivace digitálních informací není technický problém, technické problémy jsou řešitelné technickými prostředky. Větším rizikem ztráty informací v digitální podobě je nedostatek vůle a odvahy se problematikou dlouhodobé archivace začít prakticky zabývat, i přesto, že nemáme dokonalé řešení, neomezené zdroje nebo všechny odborné znalosti.

---

[42] DSpace – *http://www.dspace.org/* ; DURASPACE – *http://www.duraspace.org/* ; Islandora – *http://islandora.ca/ ; iRods – http://irods.org/*

[43] KLINDT, M. a K. AMERHEIM. *One core preservation system for all your data. No exceptions!* In: *iPRES 2015, 2-6.11.2015, Chapel Hill, USA*. Autoři článku měli k dispozici při psaní textu preprint od autora.

Použité zdroje

BROWN, Adrian. *Practical digital preservation: a how-to guide for organizations of any size*. 1st ed. London: Facet, c2013, xvi, 336 s. ISBN 978-1-85604-755-5.

CISCO. The Zettabyte Era – Trends and Analysis, White Paper. Cisco.com [online]. May 2015. [cit. 2015-07-28]. Dostupné z: *www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/VNI_Hyperconnectivity_WP.html.*

CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEMS. *Reference Model for an Open Archival Information System (OAIS): CCSDS 650.0-B-1* [online]. Washington (DC): Consultative Committee for Space Data Systems, January 2002 [cit. 2015-07-28]. 148 s.
Dostupné z: *http://public.ccsds.org/publications/archive/650x0b1.pdf.*

CORRADO, E. M., a H. L. MOULAISON. *Digital preservation for libraries, archives, and museums*. Lanham, MA: Rowman & Littlefield, 2014. 270 s. ISBN 0810887126.

FERLE, Christoph H. Marktstudie digitale Langzeitarchivierung: im Spannungsfeld zwischen Digital Preservation und Enterprise Information Archiving. Stuttgart: Fraunhofer, 2012. Dostupné také z: *http://www.swm.iao.fraunhofer.de/content/dam/swm/de/documents/Marktstudie_Digitale_Langzeitarchivierung_web.pdf.*

FOJTŮ, Andrea, Jan HUTAŘ a Marek MELICHAR. Dlouhodobá ochrana digitálních dokumentů a projekt NDK. In: *Knihovny současnosti 2011: Sborník z 19. konference, konané ve dnech 13.–15. září 2011 v Českých Budějovicích*. Ostrava: Sdružení knihoven ČR, 2011, s. 73–79. ISBN 978-80-86249-62-9. Dostupné také z: *http://www.sdruk.cz/data/xinha/sdruk/ks2011/sbornik_2011.pdf.*

GANTZ, John a David REINSEL. *The 2011 Digital Universe Study: extracting Value from Chaos* [online]. IDC, 2011 [cit. 2015-07-28]. 12 s.
Dostupné z: *http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf.*

KLINDT, M. a K. AMERHEIM. *One core preservation system for all your data. No exceptions!* In: *iPRES 2015, 2–6. 11. 2015, Chapel Hill, USA*. Autoři článku měli k dispozici při psaní textu preprint od autora.

KORB, J., & S. STRODL. Digital preservation for enterprise content: a gap-analysis between ECM and OAIS. In *7th International Conference on Preservation of Digital Objects, Vienna* (pp. 221–228). Též česky na: *https://duha.mzk.cz/clanky/dlouhodoba-ochrana-podnikovych-dokumentu-analyza-rozdilu-mezi-ecm-oais.*

LAVOIE, Brian. *The Open Archival Information System (OAIS) Reference Model: Introductory Guide (2nd Edition)*. Digital Preservation Coalition, 2014. 33 s.
Dostupné z: *http://dx.doi.org/10.7207/TWR14-02*.

LIBRARY OF CONGRESS. About – Digital Preservation (Library of Congress) [online]. [s.a.] [cit. 2015-10-03]. Dostupné z: *http://www.digitalpreservation.gov/about/.*

LIBRARY OF CONGRESS. NDSA Levels of Preservation – NDSA – Digital Preservation (Library of Congress). [online]. [s.a.] [cit. 2015-10-03].
Dostupné z: *http://www.digitalpreservation.gov/ndsa/activities/levels.html.*

MINER, M. From Theory to Action: A Pragmatic Approach to Digital Preservation Strategies and Tools. In: *2014 SAA Research Forum, Joint Annual Meeting of CoSA, NAGARA, and SAA, Washington, DC. Aug. 10–16, 2014*.
Dostupné také z: *http://powrr-wiki.lib.niu.edu/images/6/69/POWRR_outcomes.pdf.*

OWENS, T. NDSA Levels of Digital Preservation: Release Candidate One. In: The Signal: Digital Preservation [online]. Library of Congress, 2012 [cit. 2015-10-03]. Dostupné z: *http://blogs.loc.gov/digitalpreservation/2012/11/ndsa-levels-of-digital-preservation-release-candidate-one/.*

WERF, Titia van der a Bram van der WERF. *The paradox of selection in the digital age*. In: *IFLA WLIC 2014, 16-22 August 2014, Lyon, France*.
Dostupné také z: *http://library.ifla.org/1042/1/138-vanderwerf-en.pdf.*