Pilotní projekt pro low-barrier přístup
k ochraně digitálního obsahu (LTP-pilot)

Fond rozvoje CESNET č. 516R1/2014

Dlouhodobá digitální archivace

B1. Přehled a úvod do problematiky

Digitální archivace – referenční model OAIS – LTP systémy – strategie dlouhodobé ochrany

Jan Hutař, Marek Melichar

Datum: 20. 10. 2015

Obsah

1. Dlouhodobá ochrana digitálních dat – definice a hlavní problémy

Naše schopnost porozumět informačnímu obsahu digitálních objektů závisí nejen na našich znalostech ale z velké části také na technologiích, které k nalezení a použití digitálních dat potřebujeme. Zastarávání SW a HW, riziko ztráty kontextu nebo ztráty schopnosti objekty vyhledat a identifikovat a celková zranitelnost digitálních objektů jsou problémy, které je nutno řešit. V posledních letech narůstá objem generovaných digitálních dat ve všech oblastech našeho života. Některá (zdaleka ne všechna) vznikající data je třeba uchovat pro budoucí použití. Ve chvíli, kdy jde o velké objemy dat, jež jsou například výsledkem několikaletého úsilí vědeckého výzkumu či skenování fyzických předloh v paměťových institucích, kdy dokumenty už nemají fyzické předlohy vůbec nebo jsou to dokumenty kritické pro bezpečnost či zdraví, dokumenty vznikající ve státní správě atp., je naše schopnost trvale uchovávat obsah v digitální podobě velmi důležitá.

Ve zprávě *Digital Universe*, kterou každoročně vydává společnost *International Data Corporation* (IDC), je odhadována velikost existujícího digitálního prostředí (vesmíru) v počtu bitů (tedy jedniček a nul) na 1,8 zettabytů (1,8 trilionu GB) v roce 2011. Pouze třetina těchto informací má nějaký typ zálohy nebo zabezpečení, pouze polovina informací, které by měly být zabezpečeny, nějaké zabezpečení má [GANTZ a REINSEL, 2011, s. 1]. Dokážeme uložit obrovské objemy dat na velmi malých fyzických médiích, ale uložená data nedokážeme efektivně ochránit pro budoucí využití. Ještě nikdy nebyla autenticita, integrita, kompletnost dokumentů tak důležitá, jako v případě dokumentů v digitální podobě.

I pokud budeme schopni uchovat data v technologicky použitelné podobě, nemusíme být schopni porozumět jejich intelektuálnímu obsahu. Vedle ochraňovaného informačního obsahu musíme mít k dispozici další informace, které nám pomohou pochopit smysl a původ uložených informací. Bez znalosti kontextu a způsobu získání dat je soubor v UTF obsahující řádky čísel:

„2312201513.3
2412201509.8
251220156.3“

nesmyslný, pokud ale máme vysvětlující informace a informace o kontextu (měření teploty v Klementinu v poledne středoevropského času ve stupních Celsia), dávají ta samá čísla smysl.

Paměťové instituce, a nejen ony, mají za povinnost chránit a zpřístupňovat uložené dokumenty. Ochránit a zachovat digitální data v dlouhodobém horizontu je ovšem daleko těžší než např. v případě papíru. Na rozdíl od fyzických dokumentů ty digitální neexistují

v hmatatelné podobě. Před nástupem Internetu panoval názor, že ochrana digitálních dokumentů spočívá v ochraně nosiče/média, které je nese. Internet ukázal, že to byl chybný kalkul. Dnes nejsou digitální objekty vázány na žádný nosič a jejich dlouhodobá ochrana spočívá v udržení jejich vlastností, integrity, autenticity během času. Neochraňujeme nosič jako v případě fyzických dokumentů, ale informaci samotnou, tedy data, bitstream, který je někde uložen a přenášen po sítích nebo jiných fyzických médiích a informační obsah, který je v bitstreamu kódován.

Abdelaziz Abid v jednom ze svých článků připomíná, že ochrana digitálních dat je nikdy nekončící proces: "Pro ochranu analogových dokumentů většinou stačí uložení v optimálních podmínkách, dostatečná kontrola fyzického stavu a minimální využívání. U digitálních dokumentů je situace podstatně složitější. Jejich ochranu lze přirovnat k udržování ohně – je nutné se mu věnovat neustále, udržovat ho a kontrolovat. Jinak zhasne a nenávratně zmizí. Při správné péči ale může být věčný." [ABID, 2007, s. 7]

Digitální informace jsou závislé na technologii, která je uloží, dekóduje a zobrazí. I díky tomu je tyto dokumenty velmi jednoduché měnit, upravovat, aniž by to bylo patrné (při záměrné změně). Změny mohou nastat také nechtěné, např. během přenosu, opět často aniž by si toho někdo všiml. Další změny mohou nastat v době, kdy jsou data uložena na nějakém médiu. I datové nosiče, optické disky, SSD disky nebo pásky, mohou obsahovat chybné sektory nebo mohou časem ztrácet spolehlivost. Z těchto důvodů je nutné udržovat údaje v podobě metadat o provenienci, o změnách a je nutné zajistit i určitou úroveň bezpečnosti proti neoprávněným přístupům nebo nechtěným změnám, poruchám. Flexibilita digitálních informací je jejich výhodou, ale také problémem.

Problematika dlouhodobé archivace přesahuje jednotlivé země, jak akcentovala již v roce 2003 charta UNESCO „O ochraně digitálního dědictví"[1] – více viz [ABID, 2007]. Jak ve svém článku *The Digital Dark Ages? Challenges In The Preservation Of Electronic Information* napsal Terry Kuny: *„Čím déle žijeme v elektronické éře digitálních dokumentů, tím více si musíme uvědomit, že … se dostáváme do období, kdy mnohé z toho co dnes víme, co je digitálně kódováno, bude navždy ztraceno. Žijeme na počátku období digitálního temna a je na knihovnících a knihovnách, stejně jako na mniších v dávných klášterech, aby udrželi tradici vedoucí k zachování dokumentů publikovaných v současnosti.“* [KUNY, 1998] Dlouhodobá ochrana digitálních dat není luxus. Zajištění odpovídající ochrany pro digitální objekty musí být součástí instituce, jejího statutu a denní náplně stejně jako je ochrana fyzických dokumentů před vlhkostí, ohněm, krádeží apod.

[1] http://unesdoc.unesco.org/images/0013/001331/133171e.pdf#page=80

1.1 Problémy spojené s digitálními objekty

Velmi pěkný obecný úvod do problematiky zastarávání HW i SW, budoucího zpřístupnění, správy dokumentů a vůbec dlouhodobého uchování digitálních dat napsal Seamus Ross v eseji *Changing Trains at Wigan: Digital Preservation and the Future of Scholarship* [ROSS, 2000]. Člení problémy na:

- podstatou fyzické
    - stárnutí nebo poškození nosičů dat (opotřebovanost, špatné podmínky uložení, porucha infrastruktury, přírodní pohroma), případně zastarávání HW
- podstatou logické
    - lze shrnout jako možná ztráta použitelnosti, srozumitelnosti obsahu digitálních informací pro uživatele vzdálené v prostoru a čase, kterým je primárně obsah určen. Tito uživatelé nemají podporu tvůrce obsahu, který nemusí již existovat. Jediné, co bude k dispozici, jsou metadata, technická, administrativní, kontextová a jiná. SW nutný pro otevření konkrétního objektu už také nemusí být dostupný. Sami uživatelé se budou měnit, tedy jejich návyky, očekávání i technologie, které používají. Předpokladem k tomu, aby se výše uvedené nestalo, jsou metadata, která udržují veškeré nutné údaje. Logická ochrana na jejich základě umožní migrace dat a dokumentace procesů, které vedou ke zpřístupnění i v budoucích technologických prostředích.

1.2 Definice dlouhodobé ochrany digitálních dat

Dlouhodobá ochrana digitálních dat má ze své podstaty mnoho definic, které se liší úhlem pohledu. Jedna z definic „*digital preservation*“, jak zní anglický výraz, vznikla v roce 2007 v rámci pracovní skupiny Americké asociace knihoven, která publikovala tři definice, krátkou, středně dlouhou a dlouhou.

Středně dlouhá definice zní: „Dlouhodobá ochrana digitálních dat spočívá v kombinaci plánů, strategií a opatření pro zajištění přístupu k reformátovanému a digitálně vzniklému digitálnímu obsahu bez ohledu na problémy spojené s nestálostí médií a technologickými změnami. Cílem dlouhodobé ochrany digitálních dat je přesné zobrazení autentického obsahu v jakémkoliv časovém horizontu od jeho vzniku.“ [ASSOCIATION FOR LIBRARY COLLECTIONS & TECHNICAL SERVICES, 2007]

Poměrně široká definice Digital Preservation Coalition (DPC) říká, že dlouhodobá ochrana digitálních dat jsou vlastně: „všechny aktivity nutné k zajištění přístupnosti k digitálním materiálům i přes obtíže způsobené selháním médií nebo změnou technologií.“ [DIGITAL PRESERVATION COALITION, 2009c] Z pohledu jednotlivých kroků dlouhodobou ochranu specifikuje Seamus Ross: „Dlouhodobá ochrana digitálních dat má za cíl zajistit, že uživatelé v budoucnu budou moci digitální informace nalézt, získat,

zobrazit, manipulovat s nimi, interpretovat je a to vše přes konstantní změny technologií. Celý proces zahrnuje konzervaci, obnovování, výběr, mazání, vylepšování, updatování a přidávání kontextu.“ [ROSS, et al., 2009, s. 5]

Dlouhodobá ochrana je jednou ze součástí většího konceptu *digital curation* (viz např. [HARVEY, 2010]). *Digital curation* se zabývá celým životním cyklem digitálního dokumentu a směřuje k jeho dlouhodobému uchování a použitelnosti.

Ve většině projektů a plánování pro dlouhodobou ochranu digitálních dat není doba ochrany přesně specifikována. Cíle dlouhodobé ochrany jsou v realitě střednědobé s tím, že se stále průběžně upravují. Díky tomu musí být procesy ochrany kontinuální a měly by jít dále a dále do budoucnosti.

1.2.1 Aktivní (logická) a pasivní dlouhodobá ochrana digitálních dat

Stále se často setkáváme s nepochopením rozdílu mezi tzv. pasivní ochranou a aktivní logickou ochranou digitálních dat.

- Pasivní ochrana je běžná a spočívá v pouhé ochraně bitstreamu, tedy vytváření záloh a testováním použitelnosti záloh, provozování více lokací s více technologiemi apod. Ochrana bitstreamu ovšem neřeší problémy se zastaráváním SW, HW a formátů vlastních digitálních objektů, ani problémy s autenticitou. Pasivní ochrana, tedy uchování bitů, je pouze předpoklad, případně první krok, ochrany logické.
- Logická dlouhodobá ochrana digitálních dat spočívá v kontrole a procesech prováděných během životního cyklu digitálního objektu tak, aby byla zajištěna trvalá použitelnost jak objektu, tak jeho informačního obsahu. Použitelnost je obecný pojem pro vyhledatelnost, zobrazitelnost, pochopitelnost a autentičnost obsahu. Archivní objekt není navždy neměnná entita uložená na pásce nebo harddisku [FOJTŮ, HUTAŘ a MELICHAR, 2011, s. 74-75]. Naopak, k zajištění trvalé použitelnosti musí být dokumenty v archivu stále „živé“, reflektovat změny v globálním technickém prostředí a reagovat na změny, které vyžaduje správa dokumentů. Pro objekty musí být vytvořena a spravována odpovídající metadata (o vlastnostech, kontextu aj.), která se neustále doplňují a obohacují. Každá událost či změna archivního objektu má být zaznamenána v metadatech. S tím vším souvisí neustálá kontrola organizace/instituce, která ochranu zajišťuje, pomocí auditů a jiných podobných mechanismů tak, aby organizace měla např. finance na ochranu, mandát apod. Logická ochrana digitálních dat také znamená sledovat změny technologického prostředí a změny znalostí a očekávání uživatelů. Tedy je třeba mít mechanismy, které napomohou rozhodnutí o tom, kdy je třeba provést ochranné akce (*preservation actions*), nejčastěji migraci nebo emulaci nebo doplnění metadat tak, aby byl zajištěn přístup k dokumentům i v budoucnu a jejich použitelnost.

Mnoho institucí považuje dlouhodobou ochranu za pouhou zálohu dat, nevidí tu aktivní část, procesy, které umožní využití digitálních objektů v novém technologickém prostředí.

Požadavky na dlouhodobé uchování digitálních informací mohou být rozporné: Jak zajistit, aby intelektuální obsah objektů zůstal nezměněný a autentický, a byl zároveň použitelný v budoucím neznámém technologickém prostředí a byl použitelný uživateli, o jejichž znalostech nic nevíme? Obsah musí být srozumitelný skutečně nezávisle, bez přispění původního tvůrce. V situaci, kdy noví uživatelé již nemají například obecné znalosti potřebné k pochopení obsahu dokumentu, znamená to, že jsme v ochraně neuspěli? Má se archiv snažit zpětně doplnit další metadata, když se změní znalosti uživatelů? A kde je má vzít? Například archiv ukládající medicínské texty z 18tého století – měl by takový archiv ukládat vedle dat nějaké slovníky tehdejší terminologie? Měl by ukládat popis gramatiky a slovní zásoby tehdejší angličtiny? Když tenhle příklad promítneme do budoucnosti, vidíme, jak komplexní může být snaha o logickou ochranu dostupnosti a použitelnosti informačního obsahu.

2. Referenční rámec OAIS – obecný popis

Na konci 90. let 20. století rostla potřeba definice standardního rámce, který by popisoval funkční komponenty a procesy vlastní většině archivů/repozitářů určených pro dlouhodobé ukládání dat v digitální formě. Archivní komunitě chyběl obecný rámec, který by popsal společná východiska a jazyk pro další vývoj archivačních systémů. Úkolu jej vytvořit se zhostila organizace *Consultative Committee for Space Data Systems* (CCSDS), což je organizace pro mezinárodní spolupráci vesmírných agentur. Pod záštitou CCSDS vznikl v otevřeném fóru referenční model, který měl:

- ukotvit terminologii a koncepty pro popis a porovnání datových modelů a archivačních architektur;
- popsat podstatné entity a vztahy mezi nimi v archivním prostředí;
- vysvětlit klíčové funkce a informační komponenty archivačního systému.

Práce CCSDS vyústila v květnu 1999 ve vydání referenčního modelu OAIS, který byl v roce 2003 publikován jako mezinárodní norma ISO 14721:2003[2]. Poslední, doplněná verze normy, je z roku 2012.

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

[2] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=24683

- terminologický slovník pro oblast dlouhodobé ochrany a archivů/repozitářů, který je srozumitelný široké odborné veřejnosti;
- informační model a také
- funkční model digitálního archivu.
- Tedy popisuje klíčové procesy probíhající v digitálním archivu a jeho funkční komponenty[3].

2.1 OAIS balíčky a moduly

OAIS archiv pracuje v prostředí formovaném vztahy mezi čtyřmi entitami: tvůrci dat (*producers*), koncovými uživateli (*consumers*), správci (*managers*) a archivem samotným. V archivu jsou digitální objekty v podobě tzv. informačních balíčků (viz níže). Producenti dat poskytují informace (data a metadata), která jsou v archivu uložena. Uživatelé uložené informace (data a/nebo metadata) využívají. Mezi uživateli hraje významnou roli určená skupina (*designated community*), pro kterou především jsou archivované informace určeny a měla by je v budoucnu umět bez problémů najít, zobrazit a porozumět jim i jejich kontextu. Informační model OAIS pak podrobně popisuje strukturu a obsah archivního informačního balíčku (AIP), tj. říká, jaká metadata mají být spolu s uchovávaným obsahem ukládána.

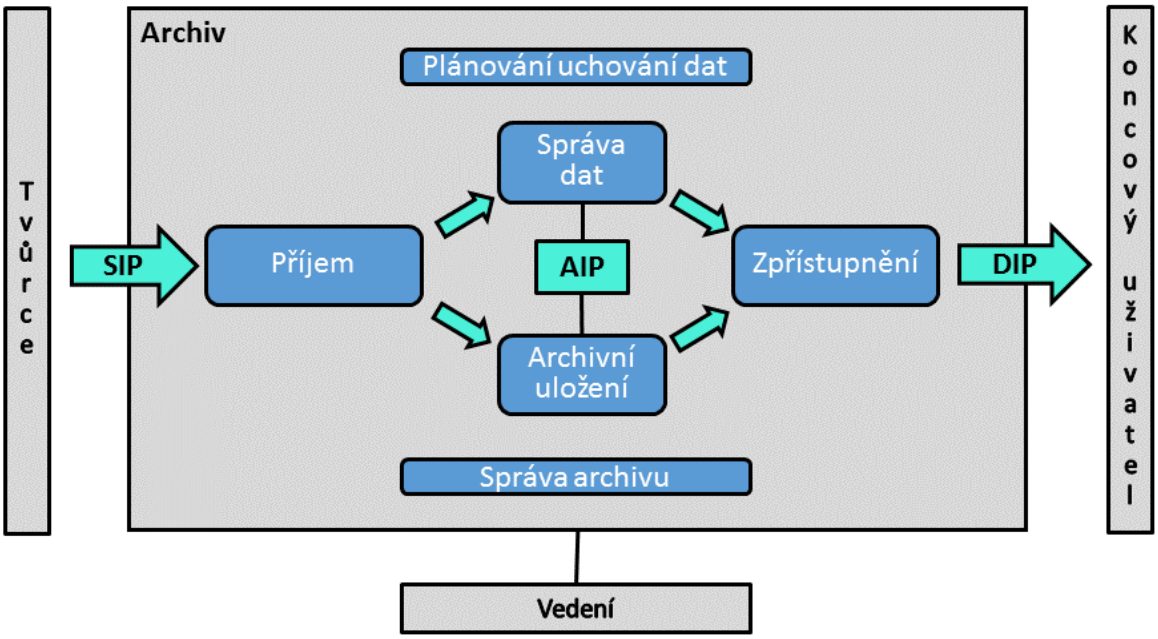
Klíčové funkce dlouhodobého archivu budovaného v souladu s koncepty OAIS, jsou popsány v šesti funkčních celcích. Tyto hlavní funkční celky ukazuje Obrázek 1 níže. Jsou to:

- ***Příjem (Ingest)*** – přijem informačního balíčku od producenta; tento celek je zodpovědný za přijetí dat a jejich přípravu na uložení a správu, tj. vytvoření balíčku AIP a doplnění potřebných metadat;
- ***Archivní uložení (Archival Storage)*** – zajišťuje, že data i jejich kontext budou uložena bezpečně; zařizuje uložení, správu a vyhledávání balíčků AIP;
- ***Správa dat (Data Management)*** – podporuje přístup k datům a jejich úpravy/správu; koordinuje popisné informace k datům spolu se systémovými informacemi používanými na podporu archivních funkcí;
- ***Správa (Administration)*** – slouží ke správě procesů, funkcí a k nastavení repozitáře i jeho uživatelů;
- ***Zpřístupnění (Access)*** – poskytuje rozhranní mezi archivem a uživatelem (obecný uživatel nebo administrátor repozitáře); pomáhá uživateli vyhledat, získat a zobrazit požadovaný archivovaný obsah repozitáře[4];

[3] Velmi dobrým textem popisujícím současný stav OAIS a jeho vliv na komunitu z roku 2014 DPC report: Brian Lavoie - The Open Archival Information System (OAIS) Reference Model: Introductory Guide (2nd Edition). Dostupné z http://dx.doi.org/10.7207/TWR14-02.

[4] Zobrazení je již mimo archiv OAIS, hlavní funkcionalita celku Zpřístupnění je dodat DIP balíček uživateli.

– ***Plánování uchovávání (Preservation Planning)*** – slouží k vytváření plánů ochrany, testování nástrojů a metod dlouhodobé ochrany a k provádění ochranných akcí.



Obrázek 1 – Referenční rámec OAIS. Dle předlohy [CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEMS, 2002, s. 4-1].

OAIS a repozitáře jemu odpovídající pracují s konceptem *informačních balíčků*. Ty obsahují jak data určená k archivaci, tak metadata o nich. Tedy balíček neobsahuje pouze data, ale také informace potřebné k ochraně digitálních objektů, které jsou jeho součástí. OAIS rozlišuje tři typy balíčků (SIP, AIP a DIP), které se od sebe samozřejmě liší a mají konkrétní úkoly v rámci repozitáře.

Vstupní informační balíček (Submission Information Package – SIP)

– je předmětem vyjednávání mezi producentem (tvůrcem informací) a OAIS archivem;
– je zasílán producentem (dodavatelem) do OAIS archivu;
– obsahuje data určená k archivaci a k nim náležející popisná, technická a případně také jiná metadata.

Archivní informační balíček (Archival Information Package – AIP)

– vzniká v OAIS archivu z balíčku SIP, nejčastěji přidáním dalších nutných metadat nebo změnou struktury SIP;
– AIP je balíček informací používaný pro archivní uložení;
– obsahuje kompletní sadu tzv. Popisné ochranné informace (Preservation Description Information – PDI) pro obsahové informace.

Výstupní informační balíček (Dissemination Information Package – DIP)

- obsahuje část nebo všechny části jednoho nebo více AIP;
- vzniká na vyžádání dat z archivu uživatelem (osoba nebo aplikace);
- OAIS je rozšiřuje uživatelům (je to vlastně to, co se dostane z repozitáře ven).

3. Možná řešení a opatření dlouhodobé ochrany digitálních dat

Digitální objekty jsou nejčastěji ohroženy ze dvou důvodů. Prvním z nich je nestálost média, na kterém jsou uloženy, druhým důvodem je závislost jejich použitelnosti a zobrazitelnosti na SW aplikaci. Strategie dlouhodobé ochrany digitálních dat můžeme rozdělit do několika skupin:

1) ochrana technologií (počítačové muzeum, ochrana HW);
2) emulace technologií (tvorba emulačního SW, případně digitální archeologie);
3) migrace informací (převod datových formátů a normalizace) a
4) ostatní (zapouzdření, přenos dat na papír, film).

Existující přístupy a metody jsou uvedeny ve stručnosti níže:

- ***Formátová migrace*** – převod digitálních objektů v konkrétním formátu do formátu jiného, použitelného na nových technologiích při zachování obsahu digitálních objektů, použitelnosti, zobrazitelnosti. Na rozdíl od obnovování, které je pouze udržováním bitstreamu (např. přesun dat na nový HW beze změny formátu dat), dojde při formátové migraci ke změně digitálního objektu, tedy původního bitstreamu. Druhem formátové migrace je normalizace – viz níže. Původní objekt se uchovává, mj. aby bylo možné se k němu v případě potřeby vrátit.
- ***Normalizace/spoléhání na standardy*** – do archivu přicházející digitální objekty se převádějí do omezeného počtu preferovaných formátů, u kterých se předpokládá, že splňují nároky na dlouhodobé uložení. Spolu s novým (normalizovaným) digitálním objektem se ukládá i původní objekt, pro budoucí použití a prokázání autenticity nebo další migrace do datových formátů v současné době neznámých. Menší počet známých a ověřených formátů je méně náročný na správu v rámci repozitáře. Repozitář ovšem musí být schopen přijmout a uložit jakýkoliv formát, pokud to bude nutné. Nesmí být omezen pouze pro formáty preferované.
- ***Obnovování nosiče dat/fyzická migrace*** – není metodou logické ochrany dat, ale ochrany základní, tedy bitstreamové. Jde o náhradu starých médií za nová shodná nebo za nová založená na nové technologii (např. přenos dat z optických disků na magnetické LTO pásky). Digitální objekt jako takový zůstává nedotčen a neměněn. Jediným problémem je zachování celistvosti a integrity. Ke kontrole integrity bitstreamu se používají kontrolní součty, vytvořené algoritmy, které využívají

bitstream digitálního objektu pro vytvoření unikátního kontrolního součtu (hashe). Tento kontrolní součet lze zpětně použít ke kontrole, zda se na původním bitstreamu něco změnilo.

- ***Emulace*** – využití aplikací napodobujících zastaralé systémy a technologické platformy na současných nebo budoucích počítačích pomocí emulace SW, HW a operačních systémů. Digitální objekt v původní podobě tak lze použít i na novém HW a SW.
- ***Digitální archeologie*** – nasazení procesů k obnovení a zpřístupnění digitálních objektů ze zastaralých nebo poškozených fyzických médií. Přístup je tedy takový, že se s daty nebude nic dělat a problém se začne řešit, až bude v budoucnu data někdo chtít použít. Problémem je, že data mohou být obnovena, ale nebude obnoven neexistující kontext a je možné, že data nebudou bez odpovídajících metadat dávat žádný smysl. Nemají žádné informace o reprezentaci ve smyslu OAIS referenčního rámce. Obnova také může trvat velmi dlouho, pokud bude vůbec úspěšná.
- ***HW muzeum/uchování technologií*** – spočívá ve shromažďování různých HW z různých technologických etap vývoje počítačů, spolu s odpovídajícím SW. Problémem je, že díky omezené životnosti technických přístrojů jakéhokoliv typu i v HW muzeu v určitém okamžiku nastane situace, kdy se rozbije poslední funkční exemplář konkrétního typu počítače, jeho součásti nebo periferie. Navíc tento přístup předpokládá, že se digitální objekty zachovají pouze na původních nosičích, což je spíše výjimka. Přístup HW muzea nepočítá ani s náročností obsluhy starých přístrojů a omezenými možnostmi zpřístupnění dokumentu.
- ***Přenos dat na klasický nosič*** – papír nebo mikrofilm. V případě papíru nejde o vhodné řešení, mnoho typů digitálních dokumentů není možné vytisknout tak, aby neztratily smysl, existující vazby nebo některé ze svých dynamických vlastností (např. webové stránky apod.). Hlavní vlastností, která by se ztratila tímto opatřením, je strojová čitelnost, a to u všech typů dokumentů, včetně těch textových. Přenos na mikrofilmy není tak bezpředmětný, jak by se mohlo zdát. Existují metody, jak zachytit na mikrofilm digitální informaci pomocí kódu ve speciální SW aplikaci. Ke čtení takového digitálního dokumentu uloženého na mikrofilmu je nutné informaci dekódovat opět pomocí konkrétního SW. Výhodou je skutečnost, že mikrofilm jako takový vydrží nejméně pět set let bez zvláštních nároků na uložení a bez zvláštních nároků na financování, oproti klasickému uložení na HW.
- ***Zapouzdření*** – zabalení digitálního objektu spolu s metadaty, identifikátory a s prvky nutnými pro pozdější přístup k němu. Kontejner tak obsahuje detaily o tom, jak má interpretovat bitstream digitálního objektu v kontejneru uloženém. Může také obsahovat informace o nárocích na emulaci, může obsahovat i SW aplikaci pro zpřístupnění, operační systém a další dokumentaci. Tento přístup je jediným přístupem ochrany digitálních dat, který se zaměřuje na ochranu digitální

informace už při jejím vzniku. Ostatní strategie ochrany se pokoušejí „přelstít“ problém zastarávání „ex-post“.

Jednotlivé možnosti řešení dlouhodobé ochrany navzájem nesoupeří, ale mohou se v jedné knihovně nebo v projektu dobře doplňovat. Dosti často data z jednoho digitálního repozitáře mohou být natolik různá, že budou aplikovány dva přístupy. Např. pro digitalizovaná data v TIFFu migrace, pro data z archivace webu to může být emulace.

Emulace a migrace jsou nejčastěji využívanými metodami logické dlouhodobé ochrany digitálních dat, jak je definuje referenční rámec OAIS.

3.1 Formátová migrace

Migraci lze obecně popsat jako přenos digitálních objektů ze starého technologického prostředí do prostředí nového.

Velkou výhodou migrace je, že jde o léty prověřený koncept, který je součástí dění okolo informačních technologií od jejich počátku. Další z výhod je, že digitální objekt je uložen v podobě, která umožní jeho okamžité využití a není třeba udržovat staré HW a SW prostředí pro tuto příležitost. Pokud je migrace zvolena jako jedna z metod dlouhodobé ochrany, musí probíhat opakovaně, vždy se budou objevovat nové a nové datové formáty, do kterých bude nutno původní digitální objekty převádět.

Nevýhodou je vyšší riziko ztráty informací a signifikantních vlastností (např. tzv. *„look and feel“*) digitálního objektu.

3.2 Emulace

Pod pojmem emulace se rozumí umělé „napodobení“ softwarového a hardwarového prostředí, které bylo typické nebo nutné pro prohlížení digitálního objektu v době jeho vzniku. Jde o jednu z metod logické dlouhodobé ochrany, která zatím stojí ve stínu migrace. Emulace může existovat na třech úrovních: na úrovni SW aplikace; na úrovni systémového SW (operační systém); a na úrovni hardwarové. V emulovaném operačním systému na emulovaném HW lze také přímo použít originální SW aplikace, pokud jsou k dispozici. Obě úrovně emulace SW jsou často proprietární, bez dostupné dokumentace a jejich emulace je proto velmi náročná. K tomu, abychom emulovali SW aplikaci nebo operační systém odpovídajícím způsobem, je nutné velmi dobře znát jejich technické vlastnosti, design a podmínky nasazení. Velmi dobrý popis emulace, jejích typů a možností, včetně typů emulátorů podávají J. van der Hoeven a H. van Wijngaarden v článku *Modular emulation as a long-term preservation strategy for digital objects* [HOEVEN a WIJNGAARDEN, 2005].

Nepopíratelnou výhodou emulace je, že můžeme mít digitální objekt v repozitáři uložený jen a pouze v jeho původní podobě. Emulace také může být u spousty typů komplexních digitálních objektů jedinou metodou, jak je zobrazit a zpřístupnit v budoucnu (webové stránky, databáze).

- Emulační software ovšem čelí stejnému problému jako digitální objekty samotné, v horizontu několika let bude sám potřebovat další emulátor, aby bylo možné původní emulátor spustit.
- Emulátory a emulovaná prostředí představují tak komplikované soustavy vztahů a závislostí, že je často problém udržovat v chodu je samotné. K úspěšné emulaci prostředí pro konkrétní digitální objekt je potřeba mít jasnou představu o prostředí, ve kterém objekt vznikl a pro které byl určen. Tyto údaje musejí být součástí tzv. ochranných metadat.
- Je nutné připojit dokumentaci k SW jak technickou, tak uživatelskou. Použití původní aplikace může být v budoucnu pro většinu uživatelů nepřekonatelný problém, včetně dohledání dokumentace.
- Právní implikace. Pěkný přehled podává článek *Legal aspects of emulation* [HOEVEN, SEPETJAN a DINDORF, 2010]. Pochybnosti o protiprávním zacházení se týkají aktivit, jako jsou přenos dat z médií chráněných proti kopírování (tedy obcházení tohoto opatření) na druhá média; použití a úpravy SW zatíženého autorskými právy; emulace patentovaných komponentů HW řešení.

Emulace tedy, i přes naděje do ní vkládané, není tak přímočarou metodou a často naráží na netušené problémy. I proto je migrace stále silně preferovaná na úkor emulace.

4. Dlouhodobá ochrana digitálních dat a různé typy institucí

Paměťové instituce se posledních dvacet let snaží o zachycení, uložení a logickou dlouhodobou ochranu digitálních dokumentů. Až donedávna nebylo možno říci, že by byly příliš úspěšné. Reálně stále hrozí ztráta digitálních informací, které nebudou nikdy podchyceny a archivovány. Technologie pro logickou dlouhodobou ochranu digitálních dat se teprve vyvíjejí a zdaleka nepředstavují hotová řešení.

Tento problém se ovšem netýká pouze paměťových institucí, možná ještě palčivější je v oblasti zdravotnictví (uchování digitálních dat o pacientech, vývoji léků, testech), průmyslové výrobě (opět uchování dokumentace, testování) a také v oblasti *e-governmentu*. Stále častěji se také řeší uchování tzv. vědeckých dat, která vznikají jako výstupy různých měření, testů, výzkumů [ROSS, 2011]. Mnohokrát jde o data, která nelze získat znovu a mají tak velikou cenu (měření klimatu, data ze satelitů apod.). U vědeckých

dat se navíc většinou jedná o obrovské objemy dat (viz hadronový urychlovač v CERN). Věda jako celek zažívá posun od experimentů prováděných v laboratoři nebo v reálném prostředí (*in vitro*) k experimentům založeným pouze na datech a jejich vyhodnocování (*in silico*). Tento trend, který se v USA nazývá někdy také *e-science* nebo *cyberscholarship*, produkuje kvanta dat a závisí na nich a na jejich uložení. Data vygenerovaná konkrétním experimentem nemohou často být vytvořena znovu (experiment nelze opakovat), je potřeba je tedy uchovat.

Pohledy na dlouhodobou ochranu v různých odvětvích i zdánlivě podobných institucí se často liší. V knihovnách se logická dlouhodobá ochrana řeší pro digitalizované dokumenty, nověji také pro born-digital dokumenty. Hlavním cílem je archivace tzv. archivní kopie (*master copy*) a hlavní důraz je na dokumentaci vzniku digitálního objektu a uložení odpovídajících metadat, která zajistí zpřístupnění a pochopení dokumentů i v budoucnu. Naproti tomu v archivech je daleko větší důraz na informace o původci, prokázání původce (elektronické podpisy), dokumentaci procesu výběru a také integritu a autenticitu. Integrita a autenticita je důležitá v dlouhodobé ochraně vždy, ovšem v archivech je naprostou podmínkou. Velmi často, na rozdíl od knihoven, jsou ukládány jednotliviny, tj. jediné kopie konkrétního dokumentu, který často má právní a jinou působnost.

5. Vývoj komplexního technického řešení – Long-term preservation (LTP) systém

Systémy na dlouhodobou logickou ochranu digitálních dat (LTP systémy), lze dělit na dvě skupiny, na systémy první a druhé generace. První generace systémů vznikala na začátku nového tisíciletí. Instituce se tehdy snažily vybudovat LTP systémy, které by odpovídaly referenčnímu rámci OAIS. Nejúspěšnější byla nizozemská národní knihovna s produktem IBM DIAS. Za podobný systém první generace můžeme považovat také DAITSS vyvíjený ve *Florida Digital Archive* používaný od roku 2005. LTP systémy druhé generace jsou již pokročilejší řešení (např. Rosetta, Preservica, Archivematica), která si berou ponaučení z pionýrských začátků. Instituce, které se účastnily vývoje a měly LTP systémy první generace (národní knihovny Nizozemí, Německa, americké univerzity), začaly po roce 2010 cítit, že jejich řešení není již dostatečné a provedly nebo plánují provést přechod na systémy druhé generace. Ty mají šanci se rozšířit daleko více než systémy předchozí, zvláště proto, že se jedná o hotová a standardně implementovatelná řešení. Bohužel, i podle průzkumu projektu PLANETS [SINCLAIR a BERNSTEIN, 2010], který proběhl s osmnácti ICT firmami, se ukazuje, že zájem vývojářských firem je malý a trh s komerčními produkty určenými pro logickou dlouhodobou ochranu digitálních dat je stále v plenkách a to přesto, že poptávka po řešeních na ochranu digitálních dokumentů

značně roste. V současné chvíli neexistuje ani náznak, že by vedle dvou komerčních systémů (Rosetta a Preservica) byl dostupný jiný nebo se alespoň vyvíjel.

Za jednu z prvních paměťových institucí, která realizovala potřebu specifikovat a vytvořit systém na logickou dlouhodobou ochranu digitálních dat, lze považovat Národní knihovnu Nizozemí (KB). Tato knihovna vytvářela od roku 1999 funkční specifikace LTP systému a stejného roku vypsala výběrové řízení na technické řešení. V uvedenou dobu nebyl na trhu žádný systém, který by měl funkcionalitu LTP systému nárokovanou KB. Výběrové řízení vyhrála firma IBM, která spolu s KB vytvořila první LTP systém vůbec. Systém byl pojmenován DIAS (*Digital Information Archiving System*) a byl implementován vedle KB také v německé národní knihovně, ovšem pouze v pilotním stavu, nikdy nebyl považován za ostrý provoz [ALTENHÖNER, 2009]. DIAS je považován za první generaci LTP systémů, kde byl jediným zástupcem. Tento systém dnešním nárokům již nevyhovuje a není dále vyvíjen ani podporován.

Podobné cíle, a to zajistit dlouhodobou ochranu svých dat, měli i v Národní knihovně Nového Zélandu (NK NZ). Impulsem bylo schválení zákona o povinném výtisku, který začal platit od roku 2003 a zahrnoval i dokumenty v digitální podobě. V roce 2004 vznikl národní program NDHA (*The National Digital Heritage Archive*) k zajištění technického řešení. V rámci projektu vznikly mezi lety 2003-2005 funkční specifikace LTP systému, který dostal jméno KRONOS a byl vyvíjen od roku 2006 ve spolupráci s firmou Endevour. Firmu Endevour později převzala izraelská firma Ex Libris, spolupráce s NK NZ pokračovala nadále. LTP systém dostal pracovní název DPS (*Digital Preservation System*) a od verze 1 se jmenuje Rosetta. Systém se dostal do operační fáze v roce 2008 a stále je vyvíjen a získává nové zákazníky[5]. Vedle Rosetty je dostupný další komerční systém, který byl poprvé implementován v Archivu britského parlamentu. Vytvořila jej britská firma Tessella pod názvem Safety Deposit Box (od roku 2014 Preservica). Je rozšířen hlavně v archivech ale i několika knihovnách po celém světě. Oba systémy jsou dle našeho názoru funkčně v podstatě rovnocenné.

Po roce 2009 se začaly objevovat open source systémy na logickou dlouhodobou ochranu digitálních dat. Velmi rozšířená je např. Archivematica nebo portugalský systém RODA. Je potřeba také uvést, že stále existují instituce, které si LTP systémy budují samy, z dostupných open source komponent, které spojují do funkčních celků. K takovýmto systémům je obtížné získat dokumentaci (např. NK Francie, KB Nizozemí).

[5] http://www.exlibrisgroup.com/category/RosettaOverview

6. Nástroje

V komunitě okolo dlouhodobé ochrany digitálních dat je od počátku velká poptávka po nástrojích, které by v určitých okamžicích životního cyklu digitálního objektu automaticky prováděly konkrétní úkoly. Mezi tyto úkoly lze zařadit tzv. charakterizaci (validace formátů, extrakce technických metadat, identifikace formátů), ochranu dat (plánování ochrany, migrace, emulace), manipulaci s daty, správu dat apod.

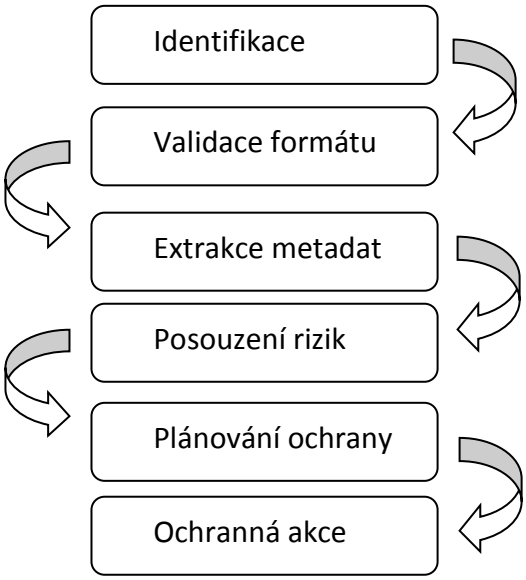
První nástroje začaly vznikat po roce 2000, jejich použitelnost a počet se od té doby zvyšuje. Nejpoužívanější jsou nástroje na charakterizaci digitálních objektů. Jde o proces získávání informací z digitálního objektu. Tyto informace jsou vytvořeny a poté uloženy ve formě metadat. Proces charakterizace má tyto části:

- *identifikace formátu* – proces ověření předpokládaného formátu digitálního objektu na základě externích příznaků (přípona názvu souboru) a vnitřních charakteristických rysů, tzv. podpisů (*format signatures*). Identifikace pomocí podpisů funguje tak, že nástroj hledá typické struktury v bitech, který tvoří soubor. Struktury mohou být jednoduché řetězce znaků nebo více řetězců v určitém pořadí na různých místech bit streamu. Každý formát má tyto struktury odlišné. Jednoduchá otázka by zněla: je soubor s příponou PDF opravdu PDF, za které se vydává? A jaká verze?
  - Nejčastěji používané nástroje jsou DROID (s pomocí databáze PRONOM), JHOVE a JHOVE2.
- *validace formátu* – proces ověření do jaké úrovně odpovídá digitální objekt syntaktickým a sémantickým pravidlům definovaných specifikací formátu. Nástroj validace prochází celý soubor, identifikuje a mapuje každou jeho sekci, porovnává míru, v jaké odpovídají specifikaci. Nakonec vytvoří výsledek, který může být podrobný, nebo jde pouze o konstatování, že soubor specifikaci, ke které se hlásí, odpovídá nebo ne. Jednoduchá otázka by zněla: je PDF validní podle specifikace PDF?
  - Nejčastěji využívané nástroje jsou JHOVE a JHOVE2, Jpylyzer.
- *extrakce vlastností objektu (metadat)* – proces získání a zaznamenání údajů o vlastnostech digitálního objektu signifikantních pro opatření dlouhodobé ochrany (tedy převážně technická metadata). Výstupem je XML záznam.
  - Nejčastěji používané nástroje jsou JHOVE a JHOVE2, NZME, Jpylyzer, mediaInfo aj.

Identifikace a validace formátů i extrakce metadat jsou nutné pro procesy dlouhodobé ochrany i pro správu digitálních objektů v digitálních repozitářích. Rozhodnutí o uložení, změnách, ochraně jsou založena nejčastěji na formátu digitálního objektu, který tedy musí být správně určen a uchován v metadatech. Existuje několik seznamů relevantních

nástrojů, zmínit bychom měli alespoň seznam knihovny Kongresu[6], registr COPTR[7] (Community Owned digital Preservation Tool Registry) nebo POWRR Tool Grid[8].

Postup zpracování digitálních objektů na vstupu do LTP systému vypadá většinou, jak ukazuje obrázek 1.



Obrázek 1 – Nejčastější podoba workflow použití nástrojů na dlouhodobou ochranu.

Podrobněji se nástrojům věnuje další text projektu LTP Pilot [B6].

[6] http://www.digitalpreservation.gov/tools/
[7] http://coptr.digipres.org/Main_Page
[8] http://digitalpowrr.niu.edu/tool-grid/

7. Přílohy

7.1 Mezinárodní spolupráce a projektová podpora

Dlouhodobá ochrana digitálních dat se velmi záhy ukázala jako natolik komplikovaný problém, že většina aktivit, výzkumů a projektů probíhá v mezinárodní spolupráci. Není v silách jedné instituce vyřešit komplexní problémy, které dlouhodobá ochrana přináší. Komunita v této oblasti zahrnuje velké i menší instituce po celém světě, jejichž odborníci jsou v podstatě v neustálém kontaktu. Spolupracují na výzkumu, vývoji nástrojů a řešení problémů pomocí emailových konferencí, běžných konferencí, sdílení nástrojů a jejich kódu např. na GitHub[9] apod.

Problémy s dlouhodobou ochranou digitálních dat se začaly objevovat v souvislosti s digitalizací a první koncepty, které ji alespoň částečně braly v potaz, se týkaly primárně digitalizace (*Parmská charta, Lundský plán* apod.). Jednou z prvních snah EU na poli dlouhodobé ochrany digitálních dat bylo zvýšení povědomí o problému, které by tak nastartovalo vlastní výzkum.

- Zástupcem prvních projektů byl evropský NEDLIB (*Networked Europian Deposit Library*) z programu *Telematics for the Libraries*, který měl za cíl navrhnout systém na dlouhodobou ochranu dle specifikace OAIS. Projekt běžel v letech 1998-2000. Přínosem bylo rozsáhlé testování emulace jako jedné z metod dlouhodobé ochrany digitálních dokumentů.
- V roce 2001 odstartoval důležitý projekt ERPANET (*Electronic Resource Preservation and Access Network*). Cílem projektu bylo vytvořit evropské konsorcium k propagaci informací, best practices a dovedností pro oblast dlouhodobé ochrany kulturních a vědeckých digitálních objektů. Za dobu trvání projektu (2001-2004) se podařilo shromáždit velké množství instruktážních materiálů, návodů, studií, pořádaly se workshopy, konference apod. Web ERPANET[10] dodnes zůstává významným informačním zdrojem, na který plynule navázal web projektu DPE (*DigitalPreservationEurope*[11]), který měl velmi podobné cíle jako ERPANET.

Začátek vlastních výzkumů a vědecké práce spočíval v definování konkrétních problémů, terminologie. Vznikly první nástroje (např. v projektech NEDLIB a DELOS), aktivity zaměřené na metadata, výběr dat, identifikaci formátů apod. Výzkum byl zaměřen převážně na „kancelářské“ dokumenty a obrazová data v institucích. Až později se řešila problematika ochrany komplexních a netradičních dokumentů. V roce 2004 volal Maurizio Lunghi, tehdy koordinátor tzv. *Florentské pracovní skupiny*, po tom, aby se dlouhodobá ochrana digitálních dat řešila koordinovaně v rámci EU, aby se vyčlenilo

[9] https://github.com/
[10] http://www.erpanet.org/
[11] http://www.digitalpreservationeurope.eu/

financování pouze na tuto oblast a aby se přešlo od výzkumu k činům [LUNGHI, 2004, s. 85-86]. Apely skupiny pro zařazení problematiky dlouhodobé ochrany digitálních dat do financování EU, se promítly do iniciativy i2010: Evropská informační společnost pro růst a zaměstnanost (*i2010: A European Information Society for Growth and Employment*), respektive *i2010: Digitální knihovny* (*i2010: Digital Libraries*) z roku 2005. Jako hlavní cíle „Iniciativy Digitální knihovny" byly stanoveny online dostupnost zdrojů, nutnost digitalizace tradičních sbírek a jejich ochrana a dlouhodobé uchovávání, což se promítlo do evropských projektů od rámcového programu 6 (FP6).

Dlouhodobé ochrany se týkaly mj. tyto projekty: FP5 (1998-2001; projekty ERPANET, DigiCULT, PRESTO, ECHO, MINERVA, FP6 (2002-2006; PRESTOSPACE, DELOS, BRICKS, CASPAR, DPE, PLANETS, PRESTOSPACE) a FP7 (2007-2013; PROTAGE, DL.ORG, LIWA, SHAMAN, KEEP, PRESTOPRIME, SCAPE, ARCOMEM, TIMBUS). Projekty z FP6, které jsou již nyní ukončeny, poskytly mnohá technická řešení a nástroje. Zvláště úspěšné ve vývoji byly projekty PLANETS[12] a CASPAR[13], nebo také KEEP[14].

Projekt PLANETS[15] (*Preservation and Long-term Access through NETworked Services*) byl více prakticky zaměřený než DPE.[16] Cílem bylo vytvoření nástrojů pro plánování dlouhodobé ochrany a testovací prostředí pro migrace a jiná opatření prováděná na digitálních objektech. Projektu se navíc účastnily firmy, které nabízejí komerční řešení pro LTP systém (Tessella, UK). Výstupy projektu jsou pro odbornou komunitu velmi podstatné. Jedná se o sadu PLANETS nástrojů, která obsahuje PLATO, PLANETS testbed, GRATE emulátor (*Global Remote Access to Emulation Services*), sada SIARD na dlouhodobou ochranu databází. Všechny nástroje jsou volně dostupné, některé jako online aplikace a jsou volně k využití. Projekt PLANETS byl velmi úspěšný a vytvořil okolo sebe komunitu odborníků, kteří chtěli spolupracovat a rozvíjet nástroje i po ukončení projektu. Po ukončení projektu proto vznikla *Open Planets Foundation*[17] (OPF, v roce 2015 přejmenována na Open Preservation Foundation) v rámci které jde vývoj dál.

Vývoj započatý v projektech PLANETS a CASPAR pokračoval v projektech SCAPE a SHAMAN (oba FP7). SCAPE měl za úkol zlepšit stav výzkumu dlouhodobé ochrany vývojem infrastruktury a nástrojů pro škálovatelnou ochranu, poskytnutím rámce pro automatická workflow. Všechny vyvinuté komponenty budou zapojeny do systému na dlouhodobou ochranu. Projekt SHAMAN vyvíjí novou generaci nástrojů pro analýzu, ingest, správu a zpřístupnění digitálních objektů. Výstupy z projektu SCAPE jsou dodnes pro komunitu velmi podstatné.

Jak vyplývá z přehledu EU výzkumu na poli dlouhodobé ochrany digitálních dat z roku 2011 [STRODL, PETROV a RAUBER, 2011], bylo na tuto problematiku vynaloženo skoro

[12] http://www.planets-project.eu/
[13] http://www.casparpreserves.eu/
[14] http://www.keep-project.eu/ezpub2/index.php
[15] http://www.planets-project.eu/
[16] Personálně i institucionálně se oba projekty prolínaly. Díky tomu měl tým NK ČR, který se účastnil DPE, velmi dobré vztahy s projektem PLANETS i s jeho řešiteli.
[17] http://openpreservation.org/

100 milionů Euro a to do více než patnácti projektů v obou rámcových programech FP6 (2002-2006) a FP7 (2007-2013). Je důležité si také uvědomit, že financování FP7 se oproti původním plánům více než ztrojnásobilo, což ukazuje na to, jaký důraz EU problému věnuje. Díky výstupům z těchto projektů máme k dispozici dnešní nástroje, znalosti a také mnoho příkladů použití v praxi.

V současnosti se aktivity na poli dlouhodobé ochrany dají shrnout pod hesla základní a aplikovaný výzkum. Základní výzkum jde za hranici jednoduchých digitálních objektů. Důraz je na interaktivních objektech, objektech vložených apod. Příkladem může být projekt LiWA (*Living Web Archives*), kterého se účastnila i NK ČR, a který se zabývá archivací dat ze sklízení webu, která jsou velmi komplexní. Podobně projekt TIMBUS, který začal v roce 2011, se věnuje výzkumu na poli ochrany business procesů.

Aplikovaný výzkum se v národních a evropských projektech věnuje vývoji škálovatelných systémů na dlouhodobou ochranu. Komunita potřebuje nástroje, metody i celé systémy k ostrému nasazení a využívání pro různorodé a velmi rozsáhlé sbírky digitálních objektů. S tím se pojí automatizace, rozhodovací procesy, výkon nástrojů, kriteria validace. V minulosti byly nástroje a moduly pro dlouhodobou ochranu navrženy tak, že vyžadovaly lidské řízení. Nyní je snaha provádět rozhodnutí i procesy automatickou cestou. Příkladem může být projekt SCAPE a také projekt ARCOMEM, který využívá sociální weby pro automatickou tvorbu informací a podporu výběru [STRODL, PETROV a RAUBER, 2011, s. 3].

Z projektů nefinancovaných z peněz EU je nutno jmenovat německý *Nestor*[18] a americký *National Digital Information Infrastructure and Preservation Program* (NDIIPP[19]), který je asi největším počinem v oblasti ochrany digitálních dat mimo EU. Dodnes fungující projekt NDIIPP začal v roce 2000, kdy si Kongresová knihovna určila za cíl vyvinout národní strategii pro ochranu digitálních informací. Od roku 2003 se začalo s projektem podle vytvořeného plánu. Od roku 2005 v rámci programu existovala speciální část věnovaná výzkumu na poli dlouhodobé ochrany. Program probíhá dodnes a to ve spolupráci s knihovnami, archivy, univerzitami v USA.

Vzhledem k tomu, že řešení dlouhodobé ochrany digitálních dat není v silách jediné instituce, vzniklo za posledních 15 let vedle projektů také několik koordinačních a podpůrných institucí.

- Ve Velké Británii existují Digital Preservatin Coalition (DPC[20]), Joint Information Systems Committee (JISC[21]), Digital Curation Centre (DCC[22]).

[18] http://www.langzeitarchivierung.de/eng/
[19] http://www.digitalpreservation.gov/
[20] http://www.dpconline.org/
[21] http://www.jisc.ac.uk/
[22] http://www.dcc.ac.uk/

- V Německu Nestor (Network of Expertise in Long-Term Storage of Digital Resources),
- v USA National Digital Information Infrastructure and Preservation Program (NDIIPP),
- v Nizozemí DANS[23] a od roku 2008 NCDD[24] (The Netherlands Coalition of Digital Preservation).

Existují také konference, jako např. iPRES[25], Archiving[26], JCDL[27], ECDL/TPDL[28] a odborné časopisy podporující sdílení znalostí v této oblasti (*International Journal of Digital Curation*[29], *Ariadne*[30], *D-Lib Magazine*[31] aj.). Nejvýznamnější z uvedených konferencí je iPRES (*International Conference on the Preservation of Digital Objects*), která se zabývá pouze a výlučně dlouhodobou ochranou digitálních dat. První ročník proběhl v roce 2004 v Pekingu. Koná se každý rok minimálně po tři dny, které jsou naplněny tutorialy, workshopy a přednáškami.

7.2 Udržení autenticity, integrity digitálních objektů

S rozmachem používání a ukládání digitálních objektů se začalo ukazovat, že digitální objekty potřebují vyšší standard autenticity a integrity než tištěné informace. Autenticita a integrita jsou důležité pro dlouhodobé uchování dat a také pro důvěryhodnost repozitáře, který musí být schopen prokázat, že to, co zpřístupňuje, je opravdu to, co bylo v repozitáři původně uloženo.

Integrita digitálního objektu (bistreamu) je ve velmi zjednodušeném pohledu dána neměnností bitů, ze kterých se skládá, případně částí, ze kterých se skládá komplexní digitální objekt. Integrita informace může ovšem spočívat také ve vlastnostech a doprovodných informacích k dokumentu se vztahujících. Zpráva *Preserving Digital Information* [GARRETT a WATERS, 1996, s. 11-18] uvádí, že integrita digitální informace má spojitost s obsahem, s celistvostí, referencemi, původem a kontextem.

- Obsahem jsou myšleny formát a struktura, které obsah vyjadřují a dovolují jej zobrazit v konkrétní aplikaci či jinak použít. Právě formát a struktura bitů působí problémy z dlouhodobého hlediska pro různé systémy. Digitální repozitář nebo LTP systém musí překonat limity vyplývající z použití konkrétního HW nebo SW

[23] http://www.dans.knaw.nl/en/content/about-dans
[24] http://www.ncdd.nl/en/index.php
[25] http://www.ifs.tuwien.ac.at/dp/ipres2010/
[26] http://www.imaging.org/ist/conferences/archiving/
[27] http://www.jcdl.org/
[28] http://www.tpdl2011.org/
[29] http://www.ijdc.net/index.php/ijdc
[30] http://www.ariadne.ac.uk/
[31] http://dlib.org/

pro čtení a zobrazení digitálních dokumentů tím, že jsou známy přesně obsah a vlastnosti těchto objektů.

- Celistvostí je myšlena integrita jednoho konkrétního diskrétního jednoduchého nebo komplexního digitálního objektu, která je nejčastěji fixována a následně validována pomocí tzv. kontrolních součtů.
- Reference souvisejí se skutečností, že digitální objekt musí mít pevnou možnost odkazování, která se nemění např. spolu s přesuny digitálního objektu. I to je jedna z věcí, které integritu dokumentu ovlivňují a dotvářejí. Aby byla zachována integrita objektu, jeho kompletnost, musí být možné jej kdykoliv přesně a bez pochyb lokalizovat mezi ostatními objekty.
- Původ (údaje o provenienci) jsou údaje o vzniku a životním cyklu digitálního objektu. Jde o prokázání všeho, co se s objektem dělo, kde a jak vznikl, tak, aby bylo možno na těchto údajích stavět procesy dlouhodobé ochrany. Koncový uživatel může v budoucnu požadovat údaje o všech změnách digitálního objektu, aby si byl jist, že se opravdu dívá na několikátou verzi originálního objektu, která se může lišit formátem i vlastnostmi, ale její obsah je autentický a s obsahem nebylo nedovoleně ani/nebo náhodně manipulováno.
- Poslední součástí, která spoludotváří integritu digitálního objektu nebo dokumentu, je kontext nebo kontextová informace. Kontext jsou vztahy mezi různými objekty nebo mezi objektem a prostředím (např. technickým i sociálním).

Poslední dvě hlediska, která ve své výše uvedené zprávě uvádějí Garrett a Waters, totiž informace o původu a kontextu, už z pohledu dnešních systémů dlouhodobé ochrany souvisejí spíše s autenticitou digitálního objektu a jejím vnímáním uživateli, producenty dat apod. Ovšem ani dnes není rozdíl mezi integritou a autenticitou digitálního objektu přijímán stejně.

Autenticita označuje nejčastěji v běžném světě skutečnost, že entita odpovídá stavu, který je pro ni výchozí nebo původní. Autenticita vždy porovnává stávající stav entity s něčím, co existovalo v minulosti. Pro digitální objekty se autenticita vztahuje nejčastěji k originálnímu objektu a jeho obsahu. Digitální objekt musí doprovázet metadata, která umožní zpětně sledovat a kontrolovat všechny procesy, které se odehrály mezi dobou, kdy byl digitální objekt vytvořen, a okamžikem, kdy si jej např. prohlíží nebo jinak používá uživatel. Z pohledu logické dlouhodobé ochrany digitálních dat je autenticita jistota, že intelektuální obsah předkládaný uživateli nebyl změněn; nebo je změněn (prošel například migrací formátu), ale o všech změnách je záznam v metadatech.

7.3 Signifikantní vlastnosti digitálních objektů a metadata

Jedna z definic říká, že signifikantní vlastnosti jsou: „Charakteristiky digitálních objektů, které musí být chráněny během doby, aby zajistily kontinuální přístupnost, použitelnost a smysl objektů.“ [WILSON, 2007] Dlouhodobá ochrana digitálních dat je o uchování

vlastností obsahu digitálních objektů skrz procesy ochrany, kterými jsou např. migrace nebo emulace. Tento proces je ale většinou takový, že ne všechny vlastnosti objektů mohou být zachovány. Je nutno si stanovit pro konkrétní sbírky, formáty dat, které vlastnosti jsou podstatné a které lze oželet. Podstatné se označují jako signifikantní vlastnosti. Ztráta vlastností nemusí proběhnout pouze po migraci, ale také díky vývoji technologií a tedy např. změnám SW aplikací, které dostatečně nepodporují předchozí verze.

Pro všechny digitální objekty je nutně potřeba jejich vlastnosti dobře zdokumentovat nejpozději v okamžiku vkládání do archivu, nejlépe v podobě metadat, která jsou uložena spolu s digitálním objektem. Ideální ovšem je, když tyto údaje provázejí objekt od okamžiku jeho vzniku. Z pohledu dlouhodobé ochrany jsou tyto údaje klíčové pro další zacházení s objektem, jeho správu a případné opatření dlouhodobé ochrany (migrace). Právě vůči těmto vlastnostem lze, a je nutné, provést srovnání výsledku migrace k původnímu obsahu. Musíme být schopni garantovat, že konkrétní nová verze digitálního obsahu je ekvivalentní ve svých signifikantních vlastnostech k objektu původnímu, i přesto, že je např. na zcela jiné technologické platformě.

Podle toho, které vlastnosti chceme zachovat, musíme vytvářet strategie ochrany, volit nástroje a specifikovat metadata a vlastně i datové formáty. Signifikantní vlastnosti jsou spojeny s autenticitou, protože tu je možno posoudit právě díky zachování a popisu vlastností konkrétního digitálního objektu. Signifikantní vlastnosti jsou dané potřebami instituce, a také zájmy cílové komunity, tedy uživatelů. U různých institucí se tedy často liší. Nejčastěji se vytvářejí přehledy signifikantních vlastností pro konkrétní datové formáty. Z technických vlastností se uvádějí a sledují mj.:

- obsah – digitální objekt je textový, obrazový,
- kontext – kdo, kdy a kde digitální objekt vytvořil,
- vzhled – font, velikost písma, barva, celkový tvar, odstavce,
- chování – hyperlinky, vzorce v MS Excel aj.,
- struktura – vložené objekty, stránkování, nadpisy.

8. Použité zdroje

ABID, Abdelaziz. 2007. Safeguarding our digital heritage: a new preservation paradigm. In: LUSENET, Yola de a Vincent WINTERMANS (eds.). *Preserving the digital heritage: principles and policies*. Amsterdam: Netherlands National Commission for UNESCO, European Commission on Preservation and Access, 2007, s. 7-14. ISBN 978-90-6984-523-4. Dostupné také z: http://www.ica.org/5697/paag-resources/preserving-the-digital-heritage-principles-and-policies.html

ALTENHÖNER, Reinhard. 2009. *Osobní rozhovor při návštěvě Národní knihovny Německa*. Frankfurt nad Mohanem, 23. duben 2009.

ASSOCIATION FOR LIBRARY COLLECTIONS & TECHNICAL SERVICES. 2007. Definitions of Digital Preservation. In: *American Library Association* [online]. Americal Library Association, 2007 [cit. 2015-07-28]. Dostupné z: http://www.ala.org/alcts/resources/preserv/defdigpres0408

CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEMS. 2002. *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

DIGITAL PRESERVATION COALITION. 2009. Introduction: Definitions and Concepts. In: *DPC Digital Preservation Handbook* [online]. Heslington, GB: Digital Preservation Coalition, 2009 [cit. 2015-07-28]. Dostupné z: http://www.dpconline.org/advice/preservationhandbook/introduction/definitions-and-concepts

FOJTŮ, Andrea, Jan HUTAŘ a Marek MELICHAR. 2011. 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. 2011. *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

GARRETT, John a Donald WATERS. 1996. *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information* [online]. The Commission on Preservation and Access and The Research Libraries Group, 1996 [cit. 2015-07-28]. 64 s. Dostupné z: http://www.clir.org/pubs/reports/pub63watersgarrett.pdf

HARVEY, Ross. 2010. *Digital curation: a how-to-do-it manual.* New York: Neal-Schuman, 2010. xxii, 225 s. How-to-do-it manuals for libraries, no. 170. ISBN 978-1-55570-694-4.

HOEVEN, Jeffrey van der a Hilde van WIJNGAARDEN. 2005. Modular emulation as a long-term preservation strategy for digital objects. In: *5th International Web Archiving Workshop (IWAW05), September 22 and 23 2005, Vienna, Austria* [online]. Alicante: European Archive, 2005 [cit. 2015-07-28]. 16 s. Dostupné z: http://iwaw.europarchive.org/05/papers/iwaw05-hoeven.pdf

HOEVEN, Jeffrey van der, Sophie SEPETJAN a Marcus DINDORF. 2010. Legal aspects of emulation. In: RAUBER, Andreas, et al. (eds.). *IPRES 2010: Proceedings of the 7th International Conference on Preservation of Digital Objects, Vienna, Austria, September 19-24, 2010.* Vienna: Oesterreichische Computer Gesellschaft, 2010, s. 113-120. ISBN 978-3-85403-262-5.

HUTAŘ, Jan. *Digitalizace, popis pomocí metadat a jejich formáty*. Praha, FF UK. 2012. Disertační práce.

KUNY, Terry. 1998. The Digital Dark Ages? Challenges In The Preservation Of Electronic Information. *International Preservation News* [online]. May 1998, no. 17 [cit. 2015-07-28]. ISSN 0890-4960. Dostupné z: http://archive.ifla.org/IV/ifla63/63kuny1.pdf

LUNGHI, Maurizio. 2004. European Actions for Sustainability and Preservation. In: Strategies for a european area of digital cultural resources: towards a continuum of digital heritage: European conference, The Hague, The Netherlands 15-16 September 2004. Haag: [s.n.], 2004, s. 79-86.

ROSS, Seamus. 2000. *Changing Trains at Wigan: Digital Preservation and the Future of Scholarship*. London: National Preservation Office, 2000. 44 s. Dostupné také z: http://www.bl.uk/aboutus/stratpolprog/collectioncare/publications/reports/wigan_digital_preservation.pdf

ROSS, Seamus. 2011. Digital Preservation: Why should today's society pay for the benefit of society in future. In: *iPRES 2011. 8th International Conference on Preservation of Digital Objects, 1.- 4. 11. 2011, Singapore.* Keynote přednáška.

SINCLAIR, Pauline a Amir BERNSTEIN. 2010. *An Emerging Market: Establishing Demand for Digital Preservation Tools and Services* [online]. PLANETS, 2010 [cit. 2015-07-28]. 10 s. Planets White Paper. Dostupné z: http://www.planets-project.eu/docs/reports/Planets-VENDOR-White-Paperv4.pdf

STRODL, Stephan, Petar PETROV a Andreas RAUBER. 2011. *Research on Digital Preservation within projects co-funded by the European Union in the ICT programme* [online]. Vienna: Vienna University of Technology, 2011 [cit. 2015-07-28]. 51 s. Dostupné z: http://cordis.europa.eu/fp7/ict/telearn-digicult/report-research-digital-preservation_en.pdf

WILSON, Andrew. 2007. *Significant Properties Report: InSPECT Work Package 2.2* [online]. V2. London: AHDS, 10/04/2007 [cit. 2015-07-28]. 10 s. Dostupné z: http://www.significantproperties.org.uk/wp22_significant_properties.pdf