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

Fond rozvoje CESNET č. 516R1/2014

Dlouhodobá digitální archivace

B6. Nástroje pro digital presevation

digital preservation – nástroje – dobrá praxe – identifikace a validace formátů – sklízení webu

Marek Melichar, Jan Hutař

Datum: 20. 10. 2015

Obsah

Úvod

V komunitě okolo dlouhodobé archivace digitálních informací 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 charakterizaci, ochranu dat (plánování ochrany, migrace, emulace), manipulaci s daty a metadaty, správu dat apod. První nástroje začaly vznikat po roce 2000, jejich použitelnost a počet se od té doby zvyšuje.

Tento text obsahuje základní informace o nástrojích, které se v praxi v oblasti dlouhodobé archivace (*digital preservation*) nebo digitálního kurátorství (*digital curation*) používají. Předem je třeba upozornit, že se nejedná o vyčerpávající přehled všech nástrojů, které jsou k dispozici. Naopak se zaměříme na nástroje, které jsou nejčastěji používané. V textu jsou také zmíněny součásti globální infrastruktury podporující praxi dlouhodobé archivace, které nemají přímo povahu nástroje, ale spíše jsou to znalostní báze a informační zdroje technického typu.

Text je určen i pro čtenáře, který se problematikou dlouhodobé ochrany digitálních dat zatím nezaobíral, který ale má zkušenosti s informačními systémy a technologiemi a je schopen experimentovat. Chtěli bychom, aby text poskytl použitelné informace, aplikovatelné v reálných situacích při správě nebo ochraně digitálních informací. Doufáme také, že by text mohl posloužit i manažerům v této oblasti nebo těm, kdo se nechtějí stát přímo specialisty na dlouhodobou archivaci nebo správu digitálních dat, ale potřebují např. vyřešit jeden konkrétní problém.

Pro úplnost je třeba říci, že nás v tomto textu především zajímají nástroje podporující logickou ochranu digitálních dat; technologiím ochrany a archivace bitového streamu se zde věnujeme jen okrajově.

Tam, kde to považujeme za užitečné, uvádíme bližší informace a techničtější podrobnosti v příkladech.

Globální informační zdroje a infrastruktura

Dlouhodobá archivace a digitální kurátorství stojí na kolektivním úsilí mnoha komunit, které poskytují své zdroje volně dalším komunitám. Žádný digitální repozitář neexistuje ve vakuu, a žádný není zcela soběstačný, každý pro svůj provoz **potřebuje informace a nástroje vytvořené jinými komunitami**. Jednoduchá řešení i komplexní systémy pro zajištění dlouhodobé

archivace a správu dat se obvykle skládají z celé řady technických a organizačních komponent.

Příklady komponent:

- HW a SW, úložná média, technické nástroje pro zpracování dat, databáze, atd.
- Lidé a jejich kompetence
- Standardy a pracovní postupy
- Finanční zdroje, organizace, mandát
- Ochraňovaná data
- Dodavatelé a uživatelé dat a jejich systémy, znalosti a potřeby
- Strategie a plány mateřské instituce
- Definice pracovních pozic
- Procesní dokumentace
- Bezpečnostní předpisy a dokumentace

Dlouhodobé uchovávání digitálních dat je **nekonečný sled rozhodnutí**, ke kterým potřebují správci digitálních dat a manažeři repozitářů odborné informace. Rozhodnutí, která mají vliv na schopnost uchovat digitální materiál, jsou jak strategické nebo manažerské povahy týkající se organizačních hledisek, tak zcela odborné a technické povahy, jako např. rozhodování o konkrétních krocích spojených s ochranou daného obrazového souboru. Odborné informace mohou správci získat z relevantní komunity, z volně dostupných zdrojů, ale také právě díky různým nástrojům a jejich výstupům. Každá lokální komunita pro dlouhodobou archivaci sdílí svoje informační zdroje a některé informační zdroje zpřístupňuje globálně, online komukoli (např. komunita uživatelů konkrétního systému pro dlouhodobou ochranu – LTP – long-term preservation systém, projekty apod.).

Příklady procesů rozhodování (více například v dokumentu ROSENTHAL, Colin a kol.[1]):

- Definice sbírek a dat, která mají byt předmětem ochrany.
- Volba HW a SW komponent (komerční vs. open source, lokální vs. SaaS – Software jako služba, pásky vs. disky apod.).
- Volba způsobu ukládání dat a jejich záloh (GRID, cloud, lokální souborové systémy).
- Návrh metod zajištění integrity dat.
- Volba a specifikace metadat, přijímaných datových formátů, struktur balíčků SIP/DIP/AIP (terminologie OAIS ISO 14721:2012).
- Stanovení kvalifikačních požadavků a organizačního uspořádání.
- Definice významných (signifikantních) vlastností ukládaného materiálu, které je potřeba dlouhodobě ochránit.
- Technická rozhodnutí o tom, co je a co není validní soubor vhodný pro archivaci.
- Strategie/pravidla na řešení problémů se soubory.

[1] ROSENTHAL, Colin a kol. Průvodce plánem důvěryhodného digitálního repozitáře (PLATTER). 1. vyd. Praha: Národní knihovna České republiky, 2009. 51 s. ISBN 978-80-7050-569-4. Online zde: http://www.ndk.cz/platter-cz

- Volba strategie ochrany (emulace, migrace, normalizace).

Příklady organizací a projektů věnujících se dlouhodobému uchovávání dat

Páteří globální komunity dlouhodobého uchovávání digitálních dat jsou organizace, které se tímto problémem přímo zabývají, sdílejí znalosti, financují výzkum a podporují např. vydávání relevantních publikací. Jde o směs organizací, které často jsou jakýmsi metodickým centrem v konkrétní zemi, nebo navazují na úspěšný projekt.

- Open Preservation Foundation (OPF) – navazuje a spravuje výsledky projektu Planets (nástroje, vzniklou komunitu, informační zdroje) a podporuje další výzkum v oblasti LTP.
- DPC (Digital Preservation Coalition), DCC (Digital Curation Centre) – sdružení podporující vzdělávání a výzkum v oblasti dlouhodobé archivace).
- Research Data Alliance (RDA) – organizace zaměřená na propagaci správy a sdílení vědeckých dat.
- POWRR (Preserving Digital Objects with Restricted Resources) – americký projekt financovaný z National Endowment for the Humanities zaměřený na provádění dlouhodobé ochrany s omezeným rozpočtem
- FADGI (Federal Agencies Guidelines Initiative) – americká organizace zaměřená na vytváření a propagaci standardů digitalizace a dlouhodobé ochrany dat
- APARSEN (Alliance for Permanent Access) – evropský projekt s množstvím užitečných výstupů (nástrojů a publikací)
- JISC – britská organizace zaměřená na využití, správu a ochranu akademických a vědeckých digitálních dat
- Národní iniciativy jako:
    - Nestor – německé sdružení pro dlouhodobou archivaci, zabývá se standardizací, podporuje výzkum, vzdělávání, audit a certifikaci).
    - NDSA (National Agenda for Digital Stewardship) – komunita institucí zabývajících se dlouhodobou archivací v USA v projektu NDIIPP a dalších projektech.
    - NCDD (Netherlands Coalition for Digital Preservation) – nizozemská komunita okolo dlouhodobé ochrany digitálních dat.
    - Digitalbevaring.dk – dánská komunita okolo dlouhodobé ochrany digitálních dat
- Velké knihovny převážně západních zemí:
    - Kongresová knihovna, USA;
    - Národní knihovna Nového Zélandu;
    - Národní archiv Velké Británie (TNA);
    - Britská knihovna;
    - Královská knihovna Nizozemí;
    - Národní knihovna Austrálie aj.

Nesmíme také zapomenout uvést, že většina amerických univerzit má long-term preservation program, např. University of Michigan a jejich ICPSR http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/preservation/index.html.

Globální informační zdroje – formátové registry

Během posledních 15 let byly klíčovými informačními zdroji pro dlouhodobou archivaci především tzv. formátové knihovny/formátové registry. Formátová knihovna PRONOM se stala hlavní součástí infrastruktury a závisí na ní řada repozitářů. Obsahuje technické informace o formátech, ale především distribuuje tzv. "signatures files", tedy soubory s definicemi, které slouží nástrojům jako je DROID, FIDO nebo Siegfried k automatické identifikaci formátů[2]. Obdobně další nástroje pro identifikaci formátů, jako je například Unix File nebo TRID mají svoje vlastní mechanismy, jak tvoří a aktualizují svoje "signatures".

Registry by měly odpovědět mj. na následující otázky (odpovídá registru PRONOM – viz dále v textu):

- Mám digitální objekt, v jakém je formátu?
- Digitální objekt uvádí, že jde o formát X, je to opravdu formát X?
- Mám objekt ve formátu X a chci jej převést na formát Y, jak?
- Mám digitální objekt ve formátu X, jaké má vlastnosti?
- Mám digitální objekt ve formátu X, jaká k němu existuje dokumentace?
- Mám digitální objekt ve formátu X, jaké je s ním spojeno riziko?
- Mám digitální objekt ve formátu X, jak a čím jej mohu zobrazit?

Z výše uvedených je nejdůležitější a nejpoužívanější registr PRONOM. Je to projekt Národního archivu Velké Británie (TNA), kde registr vznikl jako jejich interní znalostní báze spojená s nástrojem (DROID). Až později byl registr zpřístupněn celé komunitě. Jen málo lidí si ale uvědomuje, že jde stále primárně o interní projekt TNA, který se v první řadě řídí především jeho potřebami a možnostmi, nároky a potřeby komunity jsou pak sekundární. V posledních letech se rozvoj zpomalil, protože TNA nemá finance na další vývoj a doplňování obsahu. Toto je skutečná hrozba, protože většina digitálních repozitářů používá právě PRONOM/DROID na identifikaci formátů.

Pokusy nahradit PRONOM zatím neuspěly, projekt GDFR vyprodukoval například registr UDFR, kde jsou technologií sémantického webu spojeny obsahy registru formátů GDFR a PRONOM. Samotný britský národní archiv (TNA) se okolo roku 2012 pokoušel vytvořit sémantickou a lépe prohledávatelnou verzi PRONOMu, ale tato verze nebyla nikdy publikována. Kritiky PRONOMU se týkají především pomalého tempa přidávání nových informací, zastaralé

[2] Signatures jsou vlastně sekvence bitů a údaje o místě jejich výskytu v rámci bitového řetězce, které jsou typické pro konkrétní formát a jeho verzi. Nástroje, které tyto signatures používají, pomocí nich tyto formáty identifikují.

technologie a také faktu, že řada informací v PRONOMU chybí. I proto vznikají pokusy o náhradu jako např. nedávno zahájený projekt Digital preservation technical registry, který má ambice být více než jen "pouhou" formátovou knihovnou.

Nesmíme zapomenout ani na britský projekt Preserv2 (http://p2-registry.ecs.soton.ac.uk/), který je sémanticky bohatší variantou PRONOMu. Data v registru jsou z různých zdrojů, např. PRONOM, dbpedia aj.

Proč je PRONOM důležitý?

- Distribuuje soubory se "signatures" nezbytné k identifikaci formátových souborů (v podobě tzv. "signatures files releases", které pak používá nástroj DROID http://www.nationalarchives.gov.uk/aboutapps/pronom/droid-signature-files.htm).
- Udržuje a přiděluje identifikátory formátům digitálních dat (PUID), které se dnes používají v řadě komerčních a open source LTP systémech.
- PRONOM identifikátory, tzv. PUIDy, jsou do jisté míry už lingua franca v komunitě DP analytiků[3]
- Umožňuje komukoli přidat formát (tedy vlastně přidat "signature"[4]).
- XML soubor s ukázkou popisu formátu JPEG v PRONOMu.

Další relevantní alternativní zdroje k "file signatures" lze najít na těchto stránkách:

- http://www.garykessler.net/library/file_sigs.html
- http://en.wikipedia.org/wiki/List_of_file_signatures
- http://www.filesignatures.net/.

Je třeba říci, že kromě globálních a otevřených formátových knihoven existuje několik dalších komunit sdružených kolem konkrétních řešení, které společně budují vlastní formátové knihovny. Při tom využívají obvykle základní informace z PRONOMu, ale mezi členy své komunity pak sdílejí další informace o rizicích nebo pravidlech spojených se zpracováním formátů. Příklady takových komunit jsou uživatelské skupiny systémů, jako jsou Archivematica, Ex Libris Rosetta nebo SDB/Preservica.

Globální informační zdroje – znalostní báze o digitálních formátech

Kromě vlastních formátových knihoven, jejichž funkce je do značné míry technická, potřebují správci digitálního obsahu kompletní informace o digitálních formátech. Potřebují vědět, kterým digitálním formátům mohou z dlouhodobého hlediska důvěřovat a proč. Existuje

[3] Ti např. vědí, že pokud se mluví např. o fmt/353 (http://apps.nationalarchives.gov.uk/pronom/fmt/353), jde o konkrétní formát a jeho verzi, která má určité vlastnosti.

[4] Kdokoliv chce přidat do báze PRONOM nový formát, nebo udělat úpravy stávajícího, musí požádat TNA a poslat jim vysvětlení ideálně s "draftem tzv. signature" pro ten konkrétní formát. TNA se poté rozhodne, zda formát zařadí, otestuje "signature" a případně jej upraví. Celý proces často zabere několik měsíců.

několik přístupů k hodnocení vhodnosti formátu pro dlouhodobou archivaci. Prakticky největší globální dopad má přístup Kongresové knihovny, která rozlišuje sedm faktorů udržitelnosti digitálního formátu (http://www.digitalpreservation.gov/formats/). Na svém webu[5] pak udržuje rozsáhlou znalostí bázi s informace o digitálních formátech s odkazy na dokumentaci k nim, a hodnocením jejich vhodnosti pro dlouhodobé ukládání. Příklad informace o formátu JPEG2000 z databáze Kongresové knihovny: http://www.digitalpreservation.gov/formats/fdd/fdd000138.shtml

Vedle Kongresové knihovny lze nalézt i další seznamy informací o digitálních formátech:

- http://www.magicdb.org/ – starší zdroj, přestal být udržován v roce 2005
- http://en.wikipedia.org/wiki/Category:Computer_file_formats – údaje na Wikipedii
- http://www.fileformat.info/ – velmi podrobný a udržovaný zdroj informací o formátech i relevantních nástrojích
- http://fileformats.archiveteam.org/wiki/Main_Page – heslo tohoto webu je Let's Solve the File Format Problem! Obsáhlý zdroje relevantních informací.

Korpusy digitálních formátů a formátových poškození

Pro testování nástrojů na identifikaci, validaci formátů nebo i např. na extrakci metadat a na SW repozitářů je k dispozici online několik více či méně rozsáhlých korpusů digitálních objektů v různých formátech. Některé z nich také obsahují soubory s typickými chybami nebo jsou záměrně poškozené, *hacknuté nebo obfuskované*.

Pro techničtěji orientované zájemce je také online k dispozici řada nástrojů, které umožňují snadné a kontrolované poškození bitového streamu souborů nebo další podobné manipulace s bitovým streamem. Jako jednoduchý nástroj může posloužit také jakýkoli hexadecimální editor nebo textový editor.

Část informací v této oblasti pochází z komunity zabývající se forenzním výzkumem. Může být tedy zaměřena jiným směrem, než by běžný správce digitálního obsahu v paměťové instituci očekával. Přesto se vyplatí tuto komunitu neignorovat.

Velmi cenné jsou sbírky reálných poškození (glitch, color shift apod.), bohužel volně dostupných je jich jen málo.

Příklady volně dostupných formátových korpusů:

- https://github.com/openpreserve/format-corpus – korpus budovaný na webu Open Preservation Foundation, velmi rozmanitý, přispět může každý.
- http://digitalcorpora.org/corpora/govdocs – obsahuje hlavně kancelářské formáty

[5] http://www.digitalpreservation.gov/formats/fdd/browse_list.shtml

- http://www.flickr.com/groups/2121762@N23/

Globální úsilí v oblasti standardizace a certifikace důvěryhodných digitálních repozitářů

Dlouhodobá archivace především vyžaduje standardizaci (formátů dat, metadat a hlavně procesů). Základním orientačním bodem je ISO 14721 (v češtině jako ČSN ISO 14721), tzv. referenční model OAIS, který předepisuje jaké informace (metadata) mají být uloženy s ochraňovanými daty, a jaké funkce má plnit archiv odpovídající OAIS. Z OAIS jsou pak odvozeny další standardy, které například popisují mechanismy interakce mezi tvůrci obsahu a archivem (PAIMAS – ISO 20652:2006 a PAIS) nebo stanovují metodiku pro audit a certifikaci repozitářů ISO 16363 (v češtině jako ČSN ISO 16363 a dále ISO 16919). Podle OAIS je postavena architektura, datový a informační model většiny digitálních repozitářů. Uvedené normy a nástroje pro audit a certifikaci slouží především k prokázání kvality činnosti repozitáře navenek, např. pro zřizovatele, koncového uživatele apod., ale také pro vnitřní kontrolu procesů.

Vedle zmíněných ISO standardů existuje celá řada dalších nástrojů, které mohou pomoci plánování a provozování důvěryhodných dlouhodobých repozitářů. Jsou to např.:

- DRAMBORA – online nástroj na "self audit" repozitářů.
- PLATTER – nástroj na plánování budování digitálního repozitáře.
- Digital Seal of Approval (DSA) – konkrétní rámec certifikace – externě hodnocený "self audit", od roku 2014 také v českém překladu na http://dsa.cuni.cz/.
- NESTOR Criteria Catalogue – německá obdoba metodiky TRAC (a následné ISO 16363).

Za základní rámec pro audit důvěryhodného dlouhodobého digitálního repozitáře můžeme brát doporučení EU. To rozlišuje tři stupně certifikace, kterými by instituce, resp. její repozitář, měl ideálně projít postupně. Kroky se liší náročností, začíná se interním auditem a finální audit by měl být proveden externě. Jednotlivé stupně jsou uvedeny níže:

- Základní certifikace – certifikace pomocí DSA,
- Rozšířená certifikace – certifikace pomocí DSA + "self audit" podle ISO 16363 nebo DIN 31644.
- Formální certifikace – externí, nezávislý audit a certifikace dle ISO 16363 nebo DIN 31644.

V posledních dvou letech se stává velmi aktuální audit podle ISO 16363 (Audit and certification of trustworthy digital repositories). Jde o normu podporující audit a externí certifikaci digitálních repozitářů z pohledu jejich důvěryhodnosti. Norma vznikla na základě metodiky Trustworthy Repositories Audit and Certification (TRAC) a je vlastně jakýmsi

pomyslným nejvyšším formálním stupněm certifikace (viz EU metodika výše). Vyhovění tomuto standardu je důležité pro projekty financované z veřejných rozpočtů, které chtějí prokázat, že systém na digitální uchovávání, který provozují, vyhoví všem nárokům, může být považován za důvěryhodný a data jsou a budou v něm dobře ochráněna. Norma byla publikována v roce 2012, prováděcí norma ISO 16919 (Requirements for bodies providing audit and certification of candidate trustworthy digital repositories) pak v roce 2014. Existují různé pomůcky na interní audit podle této normy, zatím jen spíše tabulky pro lehčí vyplnění a vedení auditu, jako např.:

- Excelová tabulka od organizace PTAB http://www.iso16363.org/wp-content/uploads/sites/3/2014/10/Self-AssessmentTemplateforISO16363.xls.
- Nástroj založený na wiki pro provádění auditu s pomocí TRAC – PPO https://www.archivematica.org/wiki/Internal_audit_tool.

Projekty pro vzdělávání v oblasti dlouhodobé archivace a digitálního kurátorství

Vzdělávání v oblasti dlouhodobé archivace digitálních dat v Evropě bylo od počátku tisíciletí jednou z důležitých činností projektů financovaných z EU. Především projekty jako Erpanet, Interpares, nebo DigitalPreservationEurope (DPE) byly v téhle oblasti na počátku tisíciletí velmi aktivní. Od té doby se nabídka i online vzdělání, online kurzů nebo tutoriálů velmi rozšířila.

Zde uvádíme jen některé zajímavé a volně dostupné projekty:

- Online kurz Digital curation na University Colledge London https://extendstore.ucl.ac.uk/product?catalog=UCLXIDC.
- Online tutoriál Digital preservation od MIT a Cornellovy univerzity http://www.dpworkshop.org/dpm-eng/eng_index.html.
- Web projektu DPE obsahuje prezentace, články, studie v mnoha jazycích, včetně češtiny https://web.archive.org/web/*/digitalpreservationeurope.eu.
- http://www.jiscdigitalmedia.ac.uk/infokits/ – zdroj informací k různým tématům jako např. digitalizace, formáty dat apod.
- DigCurV – evropský projekt na vytvoření studijního programu a učebních osnov o dlouhodobém uchovávání http://www.digcur-education.org/.

Digital preservation nebo Digital curation jsou dnes nabízeny jako studijní obor na velkých univerzitách. V USA a Velké Británii jsou vyučovací programy zaměřené tímto směrem v posledních letech velmi časté. Knihovny těchto univerzit jsou velmi obsáhlé, a často jsou světovými centry výzkumu v oblasti dlouhodobého uchovávání digitálních dat (Harvard, MIT, University of Illinois etc.). Níže jsou uvedeny linky na informace o jednotlivých kurzech.

- King's College London https://www.kcl.ac.uk/prospectus/graduate/digital-curation
- University Aberdeen

http://www.rgu.ac.uk/information-communication-and-media/study-options/distance-and-flexible-learning/digital-curation

- University of Stuttgart – http://www.mediaconservation.org/
- University of Maine – http://digitalcuration.umaine.edu/
- Digital Curation Institute – University of Toronto http://dci.ischool.utoronto.ca/
- University of North Carolina at Chapel Hill – http://sils.unc.edu/programs/certificates/digital_curation
- Science in Information Management & Preservation (Digital) (University of Glasgow) http://www.gla.ac.uk/postgraduate/taught/informationmanagementpreservationdigitalarchivesrecordsmanagement/.

Další část se věnuje jak specializovaným nástrojům vytvořeným nebo používaným při dlouhodobém uchovávání, tak i nástrojům, které jsou užitečné při správě a manipulaci se soubory, přístupy k serverům apod.

Nástroje pro práci s digitálními objekty

Digitální soubory nebo objekty je nutno do detailu poznat, abychom je mohli dlouhodobě uchovávat. K tomu existuje mnoho nástrojů různého zaměření. Postup zpracování digitálního objektu před vstupem do dlouhodobého archivu může být sledem kroků, které mají nejprve rozpoznat, o jaký digitální formát se jedná; dále pomocí dalších nástrojů ověřit, že digitální objekt je validní reprezentací daného typu formátu. Nástroje mohou také získat technická metadata z objektu nebo z něj extrahovat důležité technické vlastnosti. Jde většinou o nástroje volně dostupné, fungující jako samostatné stojící aplikace ovládané přes příkazovou řádku, výjimečně mají nástroje uživatelské rozhraní. Velmi často jsou integrované v komplexních systémech pro dlouhodobou ochranu.

Identifikace formátů dat

Identifikace je obvykle prvním krokem při zpracování digitálního objektu v repozitáři, předchází ji kontrola "fixity", antivirové kontrole a kontrolách úplnosti balíčku dat přicházejících od producenta. Cílem vlastní identifikace digitálního formátu je nejen jednoznačné určení formátu a jeho verze, ale také získání jednoznačného identifikátoru formátu a jeho verze, které přesně určují typ souboru pro pozdější procesy. V internetovém prostředí se často používá jako identifikátor typu formátu tzv. MIME-type. Ten dokáží zjistit obvyklé nástroje operačních systémů/file systémů, ovšem pro dlouhodobou archivaci je tahle informace nedostatečná. Vedle MIME-typu totiž pracují systémy repozitářů nejčastěji s PRONOM Unique ID, tzv. PUID. K jednoznačnému určení, identifikaci formátu, používají nástroje v oblasti dlouhodobé archivace tzv. "signatures" nebo "magic numbers". Jsou to v

podstatě konkrétní bitové sekvence, které se musí nacházet v těle souboru na místech typických pro daný formát. Při identifikaci také mohou pomáhat externí znaky – jako je koncovka souboru.

Nástroje na identifikaci, resp. workflow, které tyto nástroje používá, musí řešit situace, kdy např.:

- koncovka souboru neexistuje;
- koncovka souboru neodpovídá formátu identifikovanému pomocí "signature";
- více metod identifikace nevede k stejnému výsledku (soubor s koncovkou .pdf je identifikován jako .doc);
- není k dispozici "signature", která by souborový formát uměl identifikovat (nové nebo exotické formáty);
- je identifikováno více typů/verzí formátů, mezi nimiž není nástroj schopen dále rozlišit (typicky TIFF, DOC, XML nebo TXT soubory);
- soubor je hybridní – obsahuje "signatures" více formátů (PDF a HTML/javascript v jednom, OCR TXT obsahují zdrojový kód naskenovaný k článku apod.).

Při praktickém využití nástrojů pro identifikaci velkého množství souborů musí vývojáři vždy volit mezi rychlostí a spolehlivostí. Rychlejší nástroje obvykle používají jednodušší metody identifikace (hledají sekvenci bitů v kratší části souboru nebo kontrolují jen hlavičky apod.). Větší spolehlivost znamená často vyšší nároky na čas a výpočetní výkon. Většina nástrojů tohoto typu se běžně nepoužívá jako samostatně stojící aplikace, častěji jsou zapojeny do jiné komplexní aplikace nebo systému.

Nejobvyklejší nástroje pro identifikaci souborových formátů jsou:

- DROID (Digital Record Object Identification) – http://droid.sourceforge.net/, Java, používá PRONOM "signatures" – je nástroj vyvinutý Národním archivem Velké Británie. Provádí automatickou identifikaci formátů jak u jednotlivých objektů, tak hromadně. Výstupem je informace o konkrétní verzi formátu digitálního objektu. K tvorbě výstupu využívá databázi registru formátů PRONOM, ke které je připojen. PRONOM obsahuje informace o formátech, které DROID identifikuje. Nutno podotknout, že ač JHOVE i DROID/PRONOM používají k identifikaci stejný postup (dle přípon, podpisů a sekvencí bytů), výsledky jsou často různé. Jde o známé problémy a je s nimi při práci v LTP systémech počítáno. DROID občas využívá pouze rozlišení dle přípony, pokud není schopen identifikovat formát pomocí "signatures". Použití pouze koncovky není ke kvalitní identifikaci bráno jako dostatečné. Mnoho různých formátů může mít stejnou příponu, včetně variant stejného formátu. Ve své nové verzi 6 umí DROID také kontrolovat integritu jednotlivých digitálních objektů i celých složek. V případě změny hlásí problém.
- FIDO (https://github.com/openpreserve/fido) – Python, používá PRONOM "signatures"

– velmi rychlý nástroj na identifikaci formátů digitálních objektů vyvinutý díky Open Planets Foundation komunitě. Používá podpisy z databáze PRONOM, zachází s nimi ale jinak než např. DROID, FIDO nemusí být přesné.

- Nanite (https://github.com/openplanets/nanite/) – Nanite staví na nástroji DROID a Apache Tika, takže je schopen nejen identifikovat, ale také extrahovat metadata. Je stavěn tak, aby zvládal velká množství dat, např. s použitím HADOOP.
- Siegfried (http://www.itforarchivists.com/siegfried) – Golang, používá PRONOM „signatures" – nástroj vyvinutý v roce 2014, umožňuje editaci "signatures", využívání více PRONOM "signatures" najednou, specifikaci způsobu identifikace apod.
- ffident (https://github.com/gmcgath/ffident) – java knihovna, podporuje jen některé formáty
- Unix File (http://unixhelp.ed.ac.uk/CGI/man-cgi?file=)
- Apache Tika (http://tika.apache.org/)
- TRID http://mark0.net/soft-trid-e.html

Příklad exportu metadat pro jeden soubor z nástroje DROID (výstup lze konfigurovat):

zdroj	konco vka	velikost	datum poslední úpravy	výsledek identifikace	popis formátu	verz e	mime type	PUID	metodu identifik ace
path	jp2	3.1MB	19.09.201 4 16:21	done	JP2 (JPEG 2000 part 1)		image/jp2	x-fmt/39 2	signatur e

Validace formátů a extrakce technických metadat (někdy také charakterizace)

Validace a extrakce technických metadat představují soubor kroků následující po potvrzení identity souboru. Validací se zjišťuje nakolik je daný digitální objekt v souladu s konkrétním předpisem nebo standardem pro daný typ formátu. Zjišťuje, zda jsou naplněny syntaktické požadavky (zda je objekt *well-formed*) a sémantické požadavky (zda je *valid*[6]). Charakterizace nebo extrakce technických metadat je pak další fází, jejíž cílem je získat (do metadat) v repozitáři ke každému objektu informace o pokud možno všech vlastnostech (včetně těch, které budou ochraňovány – *signifikantní vlastnosti*) digitálního objektu.[7] Každý takový nástroj pochopitelně produkuje větší množství metadat a repozitář je musí nějak konvertovat nebo

[6] Jedná se o terminologii nástroje JHOVE.

[7] Koncept *significant properties* zahrnuje jak technické vlastnosti digitálního objektu (o které jde zde) tak logické vlastnosti informačního obsahu, které nelze vždy technickými prostředky ze souboru extrahovat. I ty jsou ale předmětem zájmu dlouhodobého repozitáře. Logické *signifikantní vlastnosti* konkrétní skupiny objektů musí být popsány v dokumentaci k politikám dlouhodobé archivace a musí být jasné, co třeba uchovat v budoucnu (příklady – pořadí čtení, stránkování, barva a typ fontu, barva pozadí, kontrast textu na pozadí, uspořádání objektů na stránce, hlavička a patička atd.)

zasadit do datového modelu svého balíčku AIP.

Protože zde už se jedná o podrobnější zpracování formátů, nelze očekávat, že jeden nástroj bude umět bezchybně pracovat s jakýmkoli formátem. Podobně jako v oblasti identifikace formátů, jsou i tyto nástroje vyvíjené jako open-source a každý projekt má své limity. Nejobvyklejší nástroje, jako je JHOVE, podporují jen několik základních skupin formátů a validaci jiných neumožňují. Komunita proto našla dvě řešení – jednou cestou je vývoj formátově specifických nástrojů (jako je Jpylyzer nebo mnoho validátorů PDF/A) nebo balení více nástrojů do jednoho "wrapperu" (jako je např. FITS).

Důležité je si uvědomit, že validace znamená ověření toho, že celé tělo souboru má strukturu a obsah, které jsou stanovené nějakým předpisem, nejčastěji specifikací formátu[8]. V praxi to znamená, že k ověření validity musí nástroje analyzovat třeba i celý obsah souborů.

Samotná extrakce technických metadata nebo významných vlastností může být jednodušší. V této oblasti se běžně používají základní nástroje jako je ExifTool, který je populární i mimo oblasti dlouhodobé archivace, nebo Metadata Extraction Tool, MediaInfo a další. Ovšem tyto nástroje formáty většinou nevalidují.

Obvyklé nástroje pro validace formátů a extrakci metadat:

- JHOVE (JSTOR/Harvard Object Validation Environment) http://hul.harvard.edu/jhove/ – je velmi úspěšný a všude využívaný nástroj, vyvinutý na Harvardské univerzitě ve spolupráci s organizací JSTOR. Cílem bylo automatizovat identifikaci, validaci a extrakci metadat digitálních objektů. Nástroj pracuje s několika datovými formáty (AIFF, ASCII, BYTESTREAM, GIF, HTML, JPEG2000, JPEG, PDF, TIFF, UTF8, XML, ZIP a WAV). Je v mnoha ohledech nastavitelný, např. podoba výstupu (délka, výstupní formát txt nebo XML, obsah záznamu), způsob práce s objekty. JHOVE ale nepodporuje běžné formáty kancelářských dokumentů, kromě PDF. To může působit problémy hlavně v archivech, které tyto dokumenty budou od institucí dostávat.
- JHOVE2 https://bitbucket.org/jhove2/main/wiki/Home – pokus o následovníka JHOVE1, umí odlišné formáty, nikdy se ale masově neujal.
- NZME (New Zealand Metadata Extraction Tool) http://meta-extractor.sourceforge.net/ – jde o jeden z prvních nástrojů pro dlouhodobou ochranu vůbec. Vytvořen byl v Národní knihovně Nového Zélandu v roce 2003 se záměrem mít nástroj, který dokáže vyextrahovat ochranná metadata (převážně technická) z formátů používaných balíkem Microsoft Office. Nástroj se dodnes vyvíjí a v mnoha LTP systémech doplňuje JHOVE. Oba nástroje umí některé formáty zároveň, některé umí pouze NZME (formáty sady MS Office do verze 2007 – neumí na xml založené formáty docx, xlsx a pptx).

[8] Specifikace může být veřejná, tj. publikovaná, nebo neveřejná u proprietárních formátů (např. MS Office .doc), nebo někdy nemusí existovat vůbec.

- FITS (The File Information Tool Set) http://code.google.com/p/fits/ – jde o nástroj, který provádí identifikaci, validaci a také extrakci metadat z různých typů digitálních objektů. FITS spojuje různé obecně známé nástroje (JHOVE, ExifTool, NZME, DROID, FFident aj.) do jednoho, normalizuje a upravuje jejich výstupy a také hlásí jejich chyby. Vytvořen byl na univerzitě v Harvardu pro použití v jejich LTP systému.
- EPub validace online http://validator.idpf.org/.
- Jpylyzer https://github.com/openpreserve/jpylyzer – validace a extrakce metadat z obrazových souborů ve formátu JPEG 2000.
- ExifTool http://www.sno.phy.queensu.ca/~phil/exiftool/ – zaměřen na obrazové soubory.
- Ffprobe http://www.ffmpeg.org/ffprobe.html – mj. extrakce metadat z audio-video souborů.
- Mediainfo https://mediaarea.net/cs/MediaInfo – nástroj určený spíše pro audio a video formáty.
- BWF MetaEdit: http://bwfmetaedit.sourceforge.net/ – nástroj určený spíše pro audio a video formáty. Podporuje editaci metadat, export apod.

Seznam dalších validátorů můžete nalézt na na blogu Gary McGatha: http://www.garymcgath.com/formatsoftware.html.

Příklady výstupů několika nástrojů:

TBA sem doplnit pár xml výstupů až to bude na webu tento text

Formátová migrace (a normalizace)

Migrace digitálního objektu z jednoho formátu do jiného je jednou ze základních strategií dlouhodobé archivace. Vedle tohoto typu migrace se stejným termínem často označuje přesun dat z jednoho HW nebo replikace dat do další lokality apod. Zde ale budeme mluvit o formátové migraci, tedy převodu z jednoho formátu do druhého za účelem zajištění uchování a trvalé použitelnosti obsahu digitálního objektu, ať už proto, že formát objektu již zastarává, nebo s ním je spojeno konkrétní riziko.

Formátová migrace není jen jednoduchým převodem objektu z formátu A do formátu B.

- Správce obsahu má obvykle tisíce, miliony souborů v různých formátech. Předně musí mít každý z nich identifikovaný formát a přidělené ID formátu a jeho verze.
- Správce musí mít nástroje pro identifikaci rizikové vlastnosti (například kompresního algoritmu, který je proprietární nebo chybný) nebo musí být schopen identifikovat nevalidní soubory nebo soubory ve formátech, které nepovažuje za vhodné k archivaci. V praxi je nutné mít o každém souboru vyčerpávající technické informace, na kterých lze

později stavět. Jakmile se zjistí, že určitá vlastnost nebo technologie použitá v konkrétním formátu je riziková, je možné na základě technických metadat identifikovat postižené soubory a začít s nimi pracovat. To samé platí pro data, která jsou ve formátech, které se v určitém okamžiku formáty stanou zastaralými, ty také musí být nutné v archivu najít na základě metadat. Nevalidní soubory by měly být identifikovány již při vkládání do archivu. Pokud se správce rozhodne je uložit i přes to, že jsou nevalidní, pak musí tuto informaci uchovat v metadatech v podobě údaje o rizicích a informace o nevaliditě souboru.

- Správce musí být schopen definovat, jaké vlastnosti informačního obsahu souborů se mají při migraci zachovat. K tomu potřebuje nástroje, aby mohl možné migrační scénáře porovnat, zkontrolovat výsledky a odhadnout časové nároky migrace.
- Výsledek migrace, nový soubor a proces migrace samotný musí být popsány v metadatech objektu. Vazba mezi originálním souborem, který se zachovává a novou verzí musí být z metadat jasná a vystopovatelná.
- Pokud je repozitář skutečně budován s cílem informační obsah dlouhodobě uchovávat, měl by mít k dispozici nástroje a mechanismy, které mu umožní se o potřebě objekty migrovat dozvědět. Musí mít také dostatečně kompetentní zaměstnance, kteří dovedou s pomocí LTP systému a dalších nástrojů migraci provést a vyhodnotit rizika apod.

Z výše uvedeného je vidět, že technická realizace migrace z formátu A do formátu B je jen jedním z mnoha kroků v řetězu aktivit. Těmto aktivitám se říká “plánování dlouhodobé archivace” (preservation planning). K podpoře těchto aktivit vytvořila komunita zabývající se dlouhodobou archivací řadu metodických i praktických nástrojů (PLATO, Planets test-bed a další). Více viz kapitola Plánování dlouhodobé ochrany.

Obvyklé nástroje pro formátové migrace (používané jako samostatné aplikace nebo v podobě pluginů pro komplexní systémy dlouhodobé ochrany):

- ImageMagick: http://www.imagemagick.org/index.php Je zdaleka nejpoužívanější univerzální knihovnou v oblasti migrace obrazových formátů.
- MEncoder – http://www.mplayerhq.hu/DOCS/HTML/en/encoding-guide.html je open source nástroj na transkódování video formátů ovládaný z příkazové řádky.
- ffmpeg – http://www.ffmpeg.org/about.html podobně jako MEncoder umí přehrávat, transformovat, a streamovat video i audio formáty.
- VLC – http://www.videolan.org/vlc/index.html oblíbený přehrávač má grafické uživatelské rozhraní a umí provádět i konverze a audia.
- Jako nástroj lze použít i australský systém XENA, který je specializován na migrace formátů během ingestu (tedy normalizace) a lze jej tedy použít pouze na tento úkon – http://xena.sourceforge.net/. Je součástí australského balíku nástrojů pro podporu logické dlouhodobé ochrany DSPS (Digital Preservation Software Platform). Vyvinuto

v Národním archivu Austrálie.

- LibreOffice – http://www.libreoffice.org/ – nebo jeho knihovny jsou často používány na migrace MS Office dokumentů do PDF nebo OpenOffice formátů.
- Calibre – http://calibre-ebook.com/ – migrace různých formátů pro elektronické knihy.

Samozřejmě v určitých případech mohou být použitelné i různé další formátově specifické, a často proprietární, nástroje.

Příklad procesu plánování formátové migrace podávají McKinney a Gattuso v článku o migraci z formátu WordStar[9]. Některá komerční řešení jako Ex Libris Rosetta nebo i Preservica obsahují moduly pro plánování dlouhodobé ochrany, které poskytují "test bed" umožnující simulaci migrace pomocí více nástrojů, dále její vyhodnocení a pak teprve aplikaci na vybraná data daného typu v repozitáři.

Nástroje na práci s metadaty

Dlouhodobé uchování digitálních dat je nedílně spojeno s metadaty. Ta jsou potřebná pro ochranné aktivity, jako jsou hodnocení rizik, plánování ochranných akcí apod. V této části se budeme věnovat pouze nástrojům, které pomáhají vytvářet nebo pracovat se standardy (schématy) METS a PREMIS. Tyto dva standardy jsou často využívané v digitálních repozitářích, a proto k nim existují volně dostupné nástroje. METS a PREMIS se většinou vyskytují společně, konkrétně PREMIS je vnořen do záznamu METS. METS je metadatový standard navržený pro zápis, výměnu či sdílení různých typů metadat ke konkrétnímu digitálnímu objektu. Umožňuje zabalit do jednoho XML souboru popisná, administrativní, ochranná, strukturální metadata i metadata práv.

PREMIS na druhé straně je schéma, které je specializované na tzv. ochranná metadata, tedy metadata o digitálním objektu, která podporují procesy dlouhodobého uchovávání toho konkrétního objektu. PREMIS je jedním z doporučených standardů pro část administrativních metadat schématu METS (amdSec). Existuje konkrétní doporučení, jakým způsobem PREMIS do METS zabalit (http://www.loc.gov/standards/premis/guidelines-premismets.pdf).

Nástroje na vytváření metadat METS, PREMIS a práci s nimi

- Curator's Workbench – http://blogs.lib.unc.edu/cdr/index.php/about-the-curators-workbench/ – nástroj na vytváření METS záznamů, s vnořenými popisnými metadaty ve standardu MODS. Dokáže vytvořit METS záznam pro konkrétní objekty, s určenou strukturou apod.

[9] Converting WordStar to HTML4. Sborník iPRES 2015, s. 149-159. http://ipres2014.org/sites/default/files/upload/iPres-Proceedings-final.pdf

- PIMTOOLS – PREMIS in METS Toolbox – http://pim.fcla.edu/ – dokáže validovat PREMIS v METS záznamu, vytvořit METS záznam s PREMIS záznamem, který máme k dispozici, nebo naopak z METS záznamu, který obsahuje PREMIS kompletní PREMIS vyextrahovat.
- Sobek CM METS Editor – http://ufdc.ufl.edu/software/mets – nástroj umožňuje vytváření METS záznamu např. ze složky obsahující soubory.
- Bagit – http://www.digitalpreservation.gov/documents/bagitspec.pdf – je specifikace a také nástroj, který vytváří balíčky (bags) dle specifikace. Jako hlavní schéma používá METS. Využívá se na vytváření balíků, které je možné uložit nebo sdílet, posílat do dalšího repozitáře. Balík obsahuje manifest všech souborů, kontrolní součty apod. Nejpoužívanější je aplikace z Kongresové knihovny https://github.com/LibraryOfCongress/bagit-java.
- XFDU – XML Formatted Data Unit – je standard pro balení informací pro dlouhodobou archivaci používaný v některých komunitách využívající podobně jako METS XML (ISO 13527:2010).

Výše uvedené jsou individuální nástroje, je ale nutné říci, že vedle těchto nástrojů používá a vytváří METS a PREMIS záznamy většina systémů na správu digitálních dat (DAM – digital asset management system).

Ostatní nástroje

AVPS MDQC – http://www.avpreserve.com/tools/mdqc/ – nástroj na extrakci metadat a porovnání oproti specifickým pravidlům. Příklad použití: nástroji předložíme ukázkový soubor JP2, nástroj ukáže seznam technických metadat, na základě tohoto seznamu je možné vytvořit pravidla pro konkrétní elementy (např. "bit depth" musí být vždy 24-bit). Tato pravidla je pak možno validovat na dalších souborech a zjistit tak, zda odpovídají konkrétní specifikaci nebo mají odpovídající vlastnosti.

Speciální nástroje pro oblast archivace webu

Archivace webu (nebo Internetu) je do jisté míry specializovanou oblastí dlouhodobé ochrany digitálních dat. Cílem je stáhnout (sklízet) obsah internetových stránek a to v různém rozsahu. Národní knihovny často sklízejí obsah celé národní domény (např. doména .cz) a dělají také tematické sklizně. Instituce se mohou zaměřovat na vlastní stránky, na stránky zabývající se konkrétní tématikou apod.

- Nutným nástrojem ke sklízení je tzv. harvester (nebo crawler). Nejrozšířenější je Heritrix – https://webarchive.jira.com/wiki/display/Heritrix.
- Dále jsou nutné nástroje na indexaci sklizeného obsahu, jako např. SOLR nebo Apache Lucene.

- Pro zpřístupnění, které je v případě webového obsahu specifické, lze použít např. Wayback Machine – https://github.com/internetarchive/wayback nebo nově OpenWayback https://github.com/iipc/openwayback/wiki.
- Webové stránky jsou sklízeny a ukládány ve formátech ARC a novějším formátu WARC (ISO 28500:2009) – http://www.digitalpreservation.gov/formats/fdd/fdd000236.shtml. Pro manipulaci s těmito formáty existují různé nástroje, výše zmíněné Wayback a Heritrix s těmito formáty pracují. Další jsou např.:
    - Haward http://hawarp.openpreservation.org/
    - Tomar http://tomar.openpreservation.org/

Plánování dlouhodobé ochrany

Nejznámější volně dostupné nástroje na plánování dlouhodobé ochrany jsou výstupy projektu PLANETS, konkrétně jde o PLANETS testbed a PLATO. Jsou to aplikace na testování migrací, vytváření plánů ochrany, jejichž výstupy lze zakomponovat do workflow LTP systému. Obě aplikace mají svůj původ v evropském projektu DELOS (2002-2006), v jehož rámci existovala i sekce pro dlouhodobou ochranu digitálních dat. V této sekci vznikly první návrhy a prototypy obou nástrojů. Jejich vývoj poté pokračoval v projektu PLANETS.

- PLATO – http://www.ifs.tuwien.ac.at/dp/plato/intro/ – nástroj na plánování dlouhodobé ochrany, který umožňuje organizaci vytvářet tzv. plány ochrany, které lze automaticky provádět ve spojení s LTP systémy nebo se systémy uložení dat. Jde o online i offline prostředí, kde lze definovat požadavky na ochranu – způsoby, metody, zachování nebo pominutí signifikantních vlastností různých typů dat. Vstupem do nástroje může být manuální nastavení nebo upload předem vytvořené myšlenkové mapy, kterou si systém PLATO přeloží do jednotlivých požadavků. Nástroj také podporuje porovnání testovacích výstupů jednotlivých metod ochrany nebo výstupů z různých nástrojů, které podporuje (ImageMagick aj.). Na základě plánu ochrany je možno definovat tzv. ochranné aktivity a v systému je provádět. Např. systém EPrints je schopen od roku 2009 tyto plány ochrany pro jednotlivé typy dat (např. JPEG, GIF aj.) přímo provádět. EPrints je tak díky PLANETS doplněn o modul plánování ochrany a odpovídá více OAIS referenčnímu rámci. Stává se z něj jednoduchý LTP systém
- LTP systémy jako Rosetta, Preservica, RODA (používá PLATO) mají moduly Plánování dlouhodobé ochrany
- PLANETS Testbed – byla webová aplikace, která poskytovala prostředí na vědecké experimenty v oblasti dlouhodobé ochrany digitálních dat. Přes webový prohlížeč nabízela a kombinovala data, SW, HW nástroje k testování různých metod dlouhodobé ochrany pro různé typy digitálních objektů. Výsledky bylo možné podrobně manuálně i automaticky porovnávat. V současné době není nástroj dostupný.

Emulace

Vedle migrace byla vždy jednou ze základních strategií dlouhodobé archivace také emulace. V řadě oblastí je emulace řešením, která má vzhledem ke komplexnosti uložených dat smysl (archivace počítačových her a softwaru, archivace webu).

Emulace předpokládá, že máme k dispozici v nezměněné podobně zdrojový objekt a jsme schopni simulovat prostředí, ve kterém byl používán. To znamená, že umíme simulovat software a hardware. Nejambicióznějším pokusem o použití emulace pro účely dlouhodobé archivace byl projekt IBM nazvaný Universal Virtual Computer (UVC). Řada dalších projektu se zaměřila na zpřístupnění konkrétních sbírek dat (multimediální umění, počítačové hry apod.). Posledním větším projektem financovaným z EU fondů byl projekt KEEP, který vyprodukoval mimo jiné další verzi emulačního "frameworku" (http://emuframework.sourceforge.net/)

V praxi paměťových institucí naráží na praktické potíže s dostupností SW a ovladačů na právní potíže s licencemi v těmto komponentům.

Příklady projektů a nástrojů pro emulaci používané v oblasti dlouhodobé archivace:

- Projekt Preserving.exe
  http://www.digitalpreservation.gov/meetings/preservingsoftware2013.html?loclr=blogsig
- Emulace jako služba: http://bw-fla.uni-freiburg.de/
- Software preservation: http://www.softpres.org/
- Emulátory her a jejich prostředí: http://www.emulator-zone.com/
- Dioscuri: http://dioscuri.sourceforge.net/
- https://github.com/jsmess/jsmess
- http://jpc.sourceforge.net/home_home.html
- Emulátor PMD 85: http://pmd85.borik.net/wiki/Emul%C3%A1tor
- Informace o IQ 151: http://www.iq151.net/download.htm , http://iq151.8bit-era.cz/
- Sinclair: http://www.spectaculator.com/
- Atari: http://www.atariportal.cz/static_emulace.php, http://jpecher.sweb.cz/emulator.htm, http://karelik.wz.cz/atari.php, http://raster.infos.cz/atari/a800.htm
- Commodore: http://www.c64.cz/index.php, http://www.ccs64.com/, http://www.zzap64.co.uk/c64/c64emulators.html
- Amiga: http://www.amigaportal.cz/, http://www.amiga.cz, http://amiga.lukysoft.cz/
- Další zajímavé odkazy: http://osmi.tarbik.com/cssr/consul2717.html

Hashování a zajištění integrity

Kontrola fixity – ověření, že soubor nebyl změněn nebo poškozen, je základní operací bitové ochrany. V repozitářích probíhá obvykle na několika úrovních, v pravidelných intervalech nebo v okamžiku nějak definovaných událostí (přesuny dat, změny v infrastruktuře apod.). K ověření neměnnosti souboru se nejčastěji používají hashovací funkce nebo checksumy (MD5, SHA-2, CRC-32 apod.). Zvláštní význam má kontrola "fixity" v repozitářích, které musí prokázat, že uložené objekty jsou autentické a nezměněné z důvodu nějakých legislativních požadavků. Ty pak používají ještě další nástroje jako elektronické podpisy a časová razítka nebo třeba i kryptování obsahu.

- MD5summer – http://www.md5summer.org/ – nástroj na vytváření a validaci kontrolních součtů
- QuickHash – http://sourceforge.net/projects/quickhash/ – různé typy kontrolních součtů, výstup možno uložit jako CSV
- AVPS Fixity – http://www.avpreserve.com/tools/fixity/ – nástroj na vytváření a validaci kontrolních součtů, podporuje všemožné typy kontrolních součtů a obsahuje časový plánovač; je tedy možné nastavit pravidelné kontroly až na úroveň hodin, a to pro různé složky či lokace. Je možno také propojení s emailovým účtem, takže v případě problému dostane správce email. Fixity neukládá kontrolní součty do složky k souborům, ale do své databáze
- DROID – viz výše – vytváří kontrolní součty, pokud se tato funkcionalita nastaví a je schopen je validovat pro konkrétní složky.

Výzkumné a experimentální nástroje

V komunitě zabývající se dlouhodobou ochranou je dostupných mnoho nástrojů. Některé jsou úzce specializované, další jsou obecnější, např. na analýzu souborů v konkrétní složce. Takovým nástrojem je např. NARA File Analyzer – https://github.com/usnationalarchives/File-Analyzer a další.

Další komunitou, která vyvíjí a používá nástroje na analýzy souborů je digital forensic. Nástroje jako Bitcurator – http://www.bitcurator.net/ a zdroje jako Forensic Wiki – http://forensicswiki.org/wiki/Main_Page jsou vynikající pomocí pro procesy dlouhodobé ochrany digitální dat.

Bit level preservation

Většina tohoto textu se zabývá nástroji pro funkční nebo logickou ochranu digitálních informací. V oblasti ochrany bitového streamu se používá řada technologií a nástrojů. Ochrana bitového streamu je základem dlouhodobé archivace. Uchování vlastních bitů ale neznamená uchování informačního obsahu v nich kódovaného: bity obsahující PDF soubor mohou být za 50 let nepoužitelné, soubor obsahující informace v neznámém formátu, bez kontextu a v neznámém kódování, který se vztahuje k neznámému tématu…

Bez ohledu na použité technologie ukládání bitů je obecně možné formulovat zásady uchovávání bitů:

- uchovávání více kopií ve více lokalitách
- používání více technologií ukládání dat od více výrobců atd.
- pravidelná kontrola dat a udržování inventáře obsahu
- zálohování a testování obnov ze zálohy
- sdílení kapacit s jinými institucemi

Obecné nástroje na manipulaci s digitálními objekty

Nástroje níže jsou subjektivním výběrem autorů, nejsou v žádném případě pokusem o vyčerpávající seznam různých typů nástrojů.

Kopírování (přesun) dat

- Rsync – https://rsync.samba.org/ – je unixová aplikace na synchronizaci a transfer dat mezi dvěma lokacemi. Při transferu dat lze všemožně nastavovat. Podstatné je, že mj. kontroluje integritu dat (pomocí kontrolních součtů), zachovává vlastnosti dat (date created, modified date apod.).

Přístupy k serverům

- WinSCP – http://winscp.net/eng/docs/lang:cs – WinSCP je SFTP klient a FTP klient s otevřeným kódem (open source) pro Windows. Jeho hlavní funkcí je bezpečné přenášení souborů mezi vaším počítačem a vzdáleným serverem. Jde o jednoduchý nástroj na použití a přístup k serverům, pokud nechcete z nějakého důvodu používat přímo přes terminál.
- Putty – http://www.chiark.greenend.org.uk/~sgtatham/putty/ – terminál pro přístup k serverům.

Základní editace digitálních objektů

- PSPad – http://www.pspad.com/ – univerzální editor pro text, xml, html. Možnost vytváření maker, automatizace a nastavení kroků, pokročilá editace apod.
- Notepad++ – http://notepad-plus-plus.org/ – podobná funkcionalita jako PSPad
- LibreOffice – http://www.libreoffice.org/ – je sice kancelářský balík, je ovšem lepší pro editaci a práci např. CSV souborů, dodržuje specifikace, kódování apod, což o MS Office říci nelze.
- ExifTool – http://www.sno.phy.queensu.ca/~phil/exiftool/ – editace metadat obrazových formátů, extrakce metadat.

Odkazy na systémy a řešení pro dlouhodobou archivaci

Systémy pro dlouhodobou ochranu digitálních dat – open source:

- Archivematica: http://archivematica.org/
- Hoppla: http://www.ifs.tuwien.ac.at/dp/hoppla/
- Roda: http://roda.di.uminho.pt/?locale=en#home
- EPrints: http://preservation.eprints.org/
- DPSP: http://dpsp.sourceforge.net/ a Xena http://xena.sourceforge.net/
- Microservices: https://confluence.ucop.edu/display/Curation/Home
- Mopseus: http://194.177.192.14/mopseus/
- DAITSS: http://daitss.fcla.edu/
- Preservation Datastore: https://www.research.ibm.com/haifa/projects/storage/datastores/index.html
- Medusa + Hydra: https://wiki.duraspace.org/display/hydra/Medusa

Kompletní systémy – komerční produkty:

- Preservica Tessella: http://digital-preservation.com/solution/safety-deposit-box
- Rosetta: http://www.exlibrisgroup.com/category/RosettaOverview
- ICZ DESA/DERA: http://www.i.cz/co-delame/finance/sprava-a-rizeni-dokumentu/desa-43/
- iRods: https://www.irods.org

Existující seznamy nástrojů

- Library of Congress http://www.loc.gov/standards/mets/mets-tools.html
- http://www.loc.gov/standards/premis/tools.html
- OPF Knowledge Base:http://wiki.opf-labs.org/display/KB/Home

- OPF Tools registry: http://wiki.opf-labs.org/display/TR/Digital+Preservation+Tool+Registry
- Presto Prime, Tools: https://prestocentre.org/library/tools
- Library of Congress Tool Registry: http://www.digitalpreservation.gov/tools/
- FileFormat.Info Registry: http://www.fileformat.info/index.htm
- COPTR: http://coptr.digipres.org/Main_Page
- DCH-RP project registry http://www.digitalmeetsculture.net/heritage-showcases/dch-rp/registry-of-services-and-tools/
- POWRR Tool Grid: http://digitalpowrr.niu.edu/tool-grid/
- AVPreserve tools: http://www.avpreserve.com/avpsresources/tools/
- NARA open source doporučované nástroje – http://www.archives.gov/records-mgmt/prmd/open-source-tools-for-records-mgmt-report.pdf