Digitální knihovny

Dlouhodobé uchovávání digitálních dat

3. 4. 2024

Uchovávání digitálních dokumentů

Je třeba zajistit:

- vyhledatelnost,
- dostupnost,
- udržení srozumitelnosti,
- udržení využitelnosti,
- ochrana před ztrátou,
- ochrana před zničením,
- ochrana před ztrátou důvěryhodnosti,
- nezávislost na zastarávání nosičů digitálních dat

Digital Preservation

- Volba formátů – dat i metadat (v případě obrazových formátů TIFF vs JP2)
- Zásady – otevřený, rozšířený formát, podporovaný, adekvátní uživatelské komunitě
- Ochranná opatření – migrace, emulace
- Sledování rizik
- Důvěryhodnost, autenticita

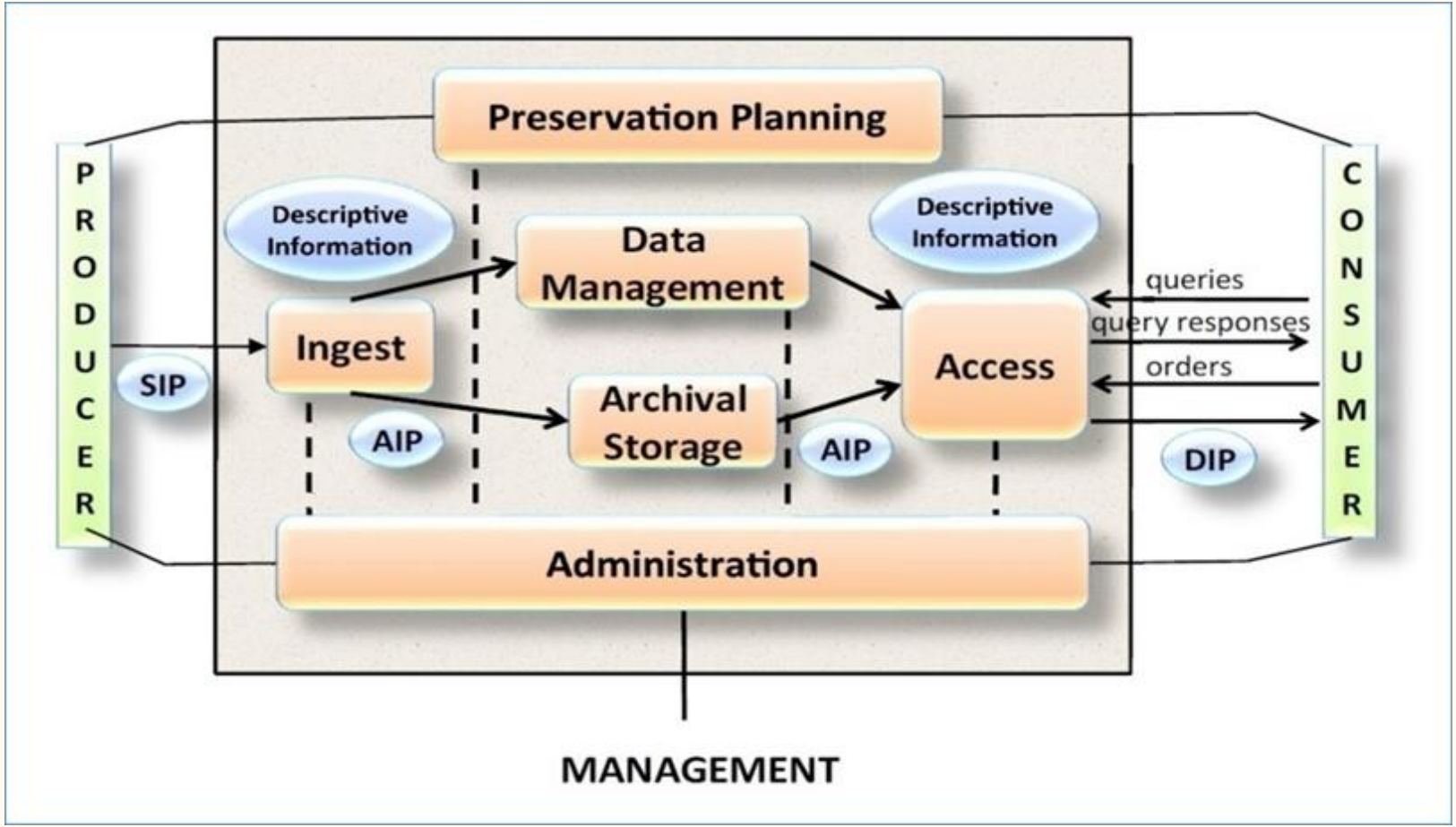
Co se může stát?

- nový upgrade softwaru nepodporuje staré typy formátů
- formát sám má novou verzi
- formát nefunguje v současné technologii

Ochrana digitálních dat

- bezpečné uložení
- bitová ochrana
- logická ochrana
- Long Term Preservation, ISO 14721 - OAIS - české názvosloví součástí normy ČSN ISO 14721 (CCSDS 650.0-M-2)
- Kolem roku 2000 první odborné skupiny a poziční dokumenty
- Koncept OAIS v roce 2000, ISO norma 2003, 2012
- SIP, AIP, DIP
- Digital Preservation Coalition a Digital Curation Centre, Alliance for Permanent Access to the Records of Science Network (APARSEN)
- 1997-2007 RLG DigiNews (první odborné periodikum), od 2006 International Journal of Digital Curation
- 2004 1. ročník iPres (dále Preservation and Archiving Special Interest Group [PASIG] a Archiving)
- World Digital Preservation Day – 1. čtvrtek v listopadu

Referenční model OAIS



Long Term Preservation System / repozitář

- Komplexní nástroj
- Archiv dle OAIS
- Repozitář, digitální archiv, LTP systém – ochrana bitstreamu i logická
- Repozitář není jen technologie sama (naopak se musí se změnami technologií vyrovnávat)

Ukládání dat dnes

- Bez ohledu na používaná média pro uložení a zálohy jde o kompromis mezi přístupností, bezpečností a náklady
- Uložení záloh na HDD – vhodné pro malé projekty
- Dnes je běžné datové úložiště pro celou instituci
  - diskové pole napojené na lokální síť
  - speciální nástroje na zálohování
    - zálohy opět na diskovém poli např. v jiném umístění, části budovy nebo alespoň 2 kopie dat
- Na správu datového úložiště je zapotřebí software
- Diskové pole – technologie RAID (Redundant Array of Inexpensive Disks)
- Erasure Coding - nová technologie, alternativa RAID, velká statická data

Digital asset management (DAM)

- je to vrstva nad hardwarem – software pro správu úložiště (dat)

- poskytuje implementaci infrastruktury k zajištění správy a ochrany dat, data a metadata v databázi, prostředí pro uložení a zpřístupnění, nástroje pro uživatele i administraci, možnost vytvoření archivu pro digitální zdroje a jejich metadata

- vrstva pro správu dle OAIS

Trvalé / dlouhodobé uchování digitálního dokumentu

- nejdříve ochrana nosiče
- snaha o hardwarová muzea, později problém s softwarem
- ochrana bitstreamu a informačního obsahu v něm
- Bitstreamová (fyzická) ochrana
  - pasivní ochrana
  - první (nezbytný) krok
  - zálohy a testováním použitelnosti záloh, provozování více lokací s víc technologiemi apod., podmínky pro provoz úložišť
  - kontrolní součty (MD5, rodina SHA)
  - základní pravidlo zálohování zní: 3-2-1. 3 kopie, 2 rozdílné technologie, 1 kopie na geograficky oddělené lokalitě

Doporučení pro bitovou ochranu

- OAIS
- National Digital Stewardship Alliance Levels of Digital Preservation
- Preservation Storage Criteria
- ISO 16363 – kapitola 5
- Zákon 499/2004 Sb. a Metodický návod č. 2/2022 odboru archivní správy a spisové služby Ministerstva vnitra pro akreditaci digitálního archivu
- Směrnice pro dlouhodobou ochranu multimediálních dat (MK ČR)
- Metodika bitové ochrany digitálních dat (projekt ARCLib)
- Pečeť Nestoru

Postupy bitové ochrany

- pravidlo 3-2-1
- pravidelné kontroly integrity (cca za 3 roky celé úložiště)
- údržba hardware, včasná výměna
- monitoring zařízení
- vhodné úložné technologie
- opatření proti lidskému faktoru (WORM - Write once read many - média)
- uchovávat historické verze dat
- mít dlouhodobý plán udržitelného provozu (organizační postupy, lidské zdroje, uchování a předávání znalostí, udržitelné financování, plánovaná pravidelná obnova hardware a software)

Postupy logické ochrany

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).

Nelze zajistit jen technickými nástroji, ty jsou však nezbytné.

Content data object

- Ochrana informačního obsahu
- CDO – Content data object
- Interpretační informace (representation information)
- Jeden objekt CDO – jeden či více souborů
- Př. Jeden objekt CDO (filmový snímek) – tři komponenty (zvuková / zvukový kodek, obrazový / obrazový kodek, strukturální / kontejnerový formát)

Data a metadata

- Datový obsah
- Popisná metadata
- Metadata o struktuře balíčku
- Administrativní metadata
- Identifikátory - uuid, URN:NBN

Souborové formáty pro archivaci - kritéria

- Ideálně přijímat do archivu omezené množství datových formátů

- Identifikace pomocí PUID (PRONOM) - formátový registr, snahy o další, aktuálně FDD
- Méně vhodné pomocí MIME type (původně pro mailovou komunikaci, 7 typů a doplněné subtypy – např. image/gif nebo application/json).

- Vhodné zvolit formát již při vzniku dokumentu

JPEG - PUID

PRONOM Unique ID	Format Name	Format Version	Extension	Format Risk
fmt/41	Raw JPEG Stream		jpe jpg jpeg	
fmt/42	JPEG File Interchange Format	1.00	jpeg jpe jpg	
fmt/43	JPEG File Interchange Format	1.01	jpg jpe jpeg	
fmt/44	JPEG File Interchange Format	1.02	jpg jpe jpeg	
x-fmt/398	Exchangeable Image File Format (Compressed)	2.0	jpg jpeg	
x-fmt/390	Exchangeable Image File Format (Compressed)	2.1	jpg jpeg	
x-fmt/391	Exchangeable Image File Format (Compressed)	2.2	jpg jpeg	
fmt/645	Exchangeable Image File Format (Compressed)	2.2.1	jpg jpeg	
fmt/1507	Exchangeable Image File Format (Compressed)	2.3.x	jpg jpeg	
fmt/112	Still Picture Interchange File Format	1.0	spf jpg	

Kritéria pro výběr vhodného formátu dat

- Nejpropracovanější v oblasti obrazových formátů
- Jak hodnotit zastaralost a ohrožení? Znáte zastaralé formáty?
- Technická kvalita (ztrátová / beztrátová komprese)
- Velikost, účel použití (velikost omezující ano/ne)
- Uživatelská komunita
- Rozšíření, podpora, licence, dokumentace

Doporučení

ii. Photographs - Digital

ii. Photographs - Digital	Preferred	Acceptable
A. Faithful representation of the work	› Equal in quality to the published version, best edition or master copy › In the same format as the master copy	
B. Technical Characteristics	› Highest resolution available, not rescaled or interpolated › Highest bit depth available, 16 bits per channel if available › Embedded color profile or specified color space used in published version › Uncompressed › Unlayered	› Lossless compression or lower compression ratios › Discrete wavelet transform (DWT) preferred to discrete cosine transform (DCT) › Layered, if supported by preferred or acceptable format
C. Formats	› TIFF (*.tif) › JPEG2000 (*.jp2) › PNG (*.png) › JPEG/JFIF (*.jpg)	› Photoshop (*.psd) › JPEG2000 Part 2 (*.jpf, *.jpx) › Digital Negative DNG (*.dng) › Proprietary Camera Raw formats (*.nef, *.crw) › GIF (*.gif)

- Doporučení Národní knihovny ČR
- JPEG2000, TIFF,
- EPUB
- PDF/A 1, 2, 3
- *.wav, *.mp4 (kodek MPEG-4 Media File)
- Dle archivní vyhlášky: PDF/A, PNG, JPEG, TIFF, GIF, MP2, MP3, MPEG-1, MPEG-2, MPEG-4, WAV, XML, DTD

Nástroje

Data

- produkce
- identifikace
- charakterizace
- validace
- prezentace
- uložení a správa

Metadata

- produkce / editace
- validace
- prezentace / interpretace
- uložení a správa

Nástroje kontroly integrity

Manuální

- Double Commander
- File Checksum Tool
- Bitrot Detector

Pokročilé

- politikami řízené úložiště

Formátový registr Pronom

PUID

Signatures

Vazba na DROID, ale i Siegfried

MagicNumber - specifické bitové sekvence

Identifikace

- DROID (https://github.com/digital-preservation/droid)
- FIDO (https://github.com/openpreserve/fido)
- Siegfried (https://github.com/richardlehane/siegfried) – nadstavba Brunhilda
- Apache Tika (http://tika.apache.org/)
- Nanite (https://github.com/openplanets/nanite/)

Validace (data i metadata)

- Jpylyzer – JPEG2000 (https://jpylyzer.openpreservation.org/)
- VeraPDF – PDF/A (https://verapdf.org/home/)
- MediaConch – zejména Matroska (https://mediaarea.net/MediaConch)
- Komplexní validátor NDK (https://www.ndk.cz/archivace/komplexni-validator)
- Validátor ZAF – SIP balíčky ESSS (https://validatorzaf.github.io/zaf/)
- Validátor NDA
- DPF manager (http://dpfmanager.org/)
- EpubCheck (https://github.com/w3c/epubcheck/releases)
- JHOVE
- FITS - wrapper (obsahuje: ADL Tool, Apache Tika, DROID, Exiftool, Jhove, MediaInfo)

JHOVE

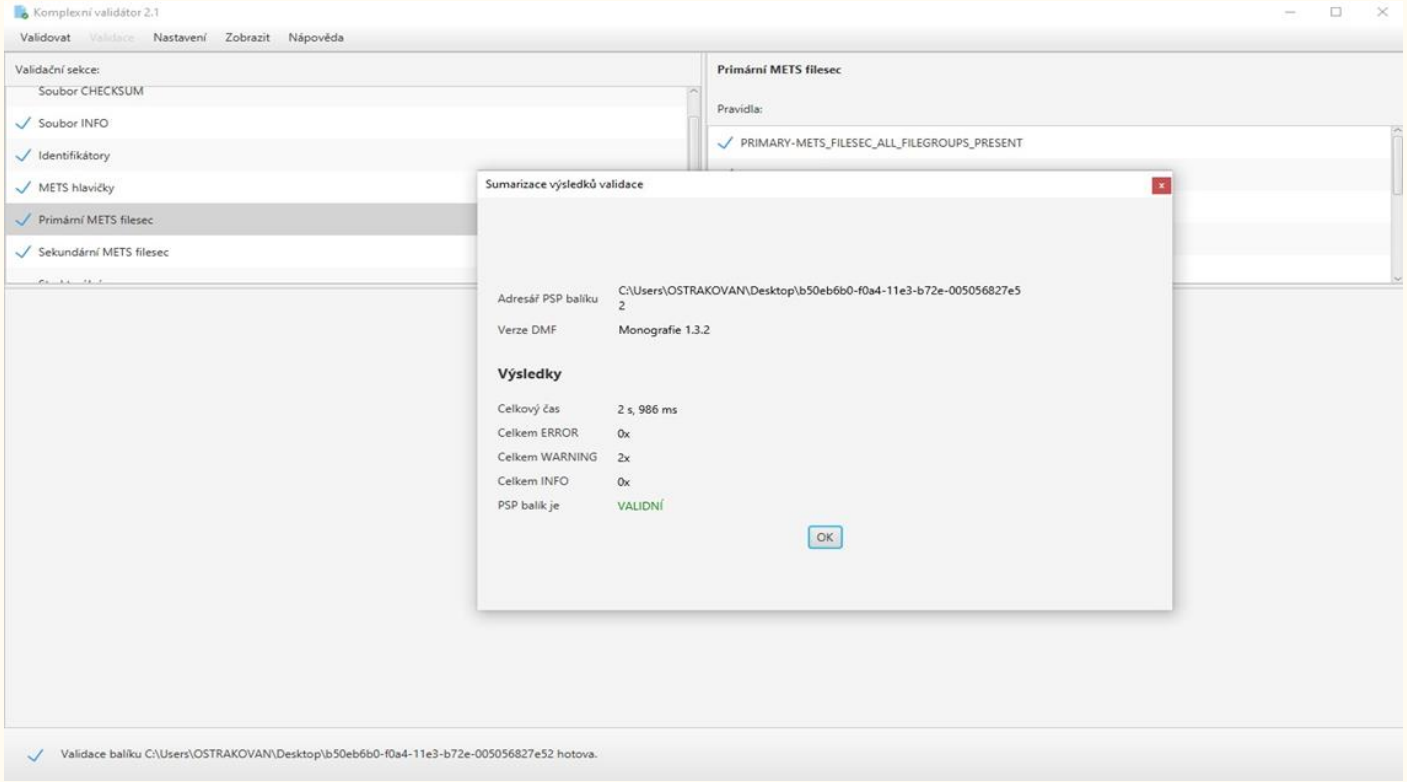
https://jhove.openpreservation.org/

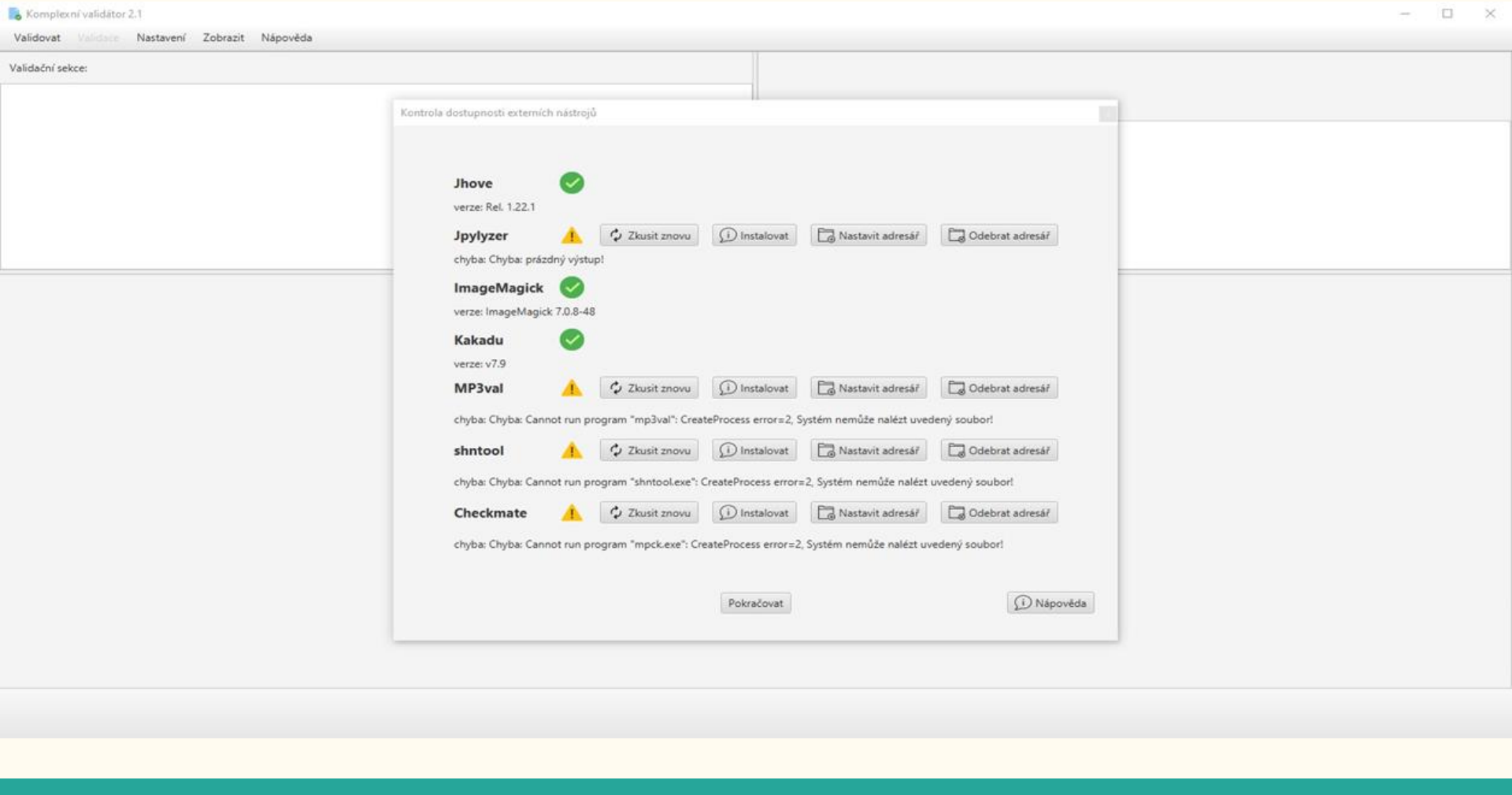
Harvard Library, JSTOR

GIF, HTML, JPEG, JPEG 2000, PDF, TIFF, WAV, XML

Komplexní validátor NDK

https://old.ndk.cz/archivace/komplexni-validator




Komplexní validátor 2.1
Validovat
Validace
Nastavení
Zobrazit
Nápověda
Validační sekce:
Kontrola dostupnosti externích nástrojů
Jhove
verze: Rel. 1.22.1
Jpylyzer
Zkusit znovu
Instalovat
Nastavit adresář
Odebrat adresář
chyba: Chyba: prázdný výstup!
ImageMagick
verze: ImageMagick 7.0.8-48
Kakadu
verze: v7.9
MP3val
Zkusit znovu
Instalovat
Nastavit adresář
Odebrat adresář
chyba: Chyba: Cannot run program "mp3val": CreateProcess error=2, Systém nemůže nalézt uvedený soubor!
shntool
Zkusit znovu
Instalovat
Nastavit adresář
Odebrat adresář
chyba: Chyba: Cannot run program "shntool.exe": CreateProcess error=2, Systém nemůže nalézt uvedený soubor!
Checkmate
Zkusit znovu
Instalovat
Nastavit adresář
Odebrat adresář
chyba: Chyba: Cannot run program "mpck.exe": CreateProcess error=2, Systém nemůže nalézt uvedený soubor!
Pokračovat
Nápověda

Charakterizace / extrakce metadat

- charakterizace / extrakce technických metadat
  - další fáze, cílem je získat (do metadat) v repozitáři ke každému objektu informace o všech vlastnostech (ochraňovány budou signifikantní vlastnosti) digitálního objektu
    - vzniká větší množství metadat: repozitář je konvertuje a ukládá do AIP
    - různé formáty – různé nástroje
    - JHOVE: podpora několika základních skupin formátů,
    - FITS – wrapper (obsahuje: ADL Tool, Apache Tika, DROID, Exiftool, Jhove, MediaInfo,
    - NZME - National Library of New Zealand Metadata Extractor,
    - EpubCheck
    - ExifTool
    - Mediainfo https://mediaarea.net/cs/MediaInfo – nástroj určený spíše pro audio a video formáty

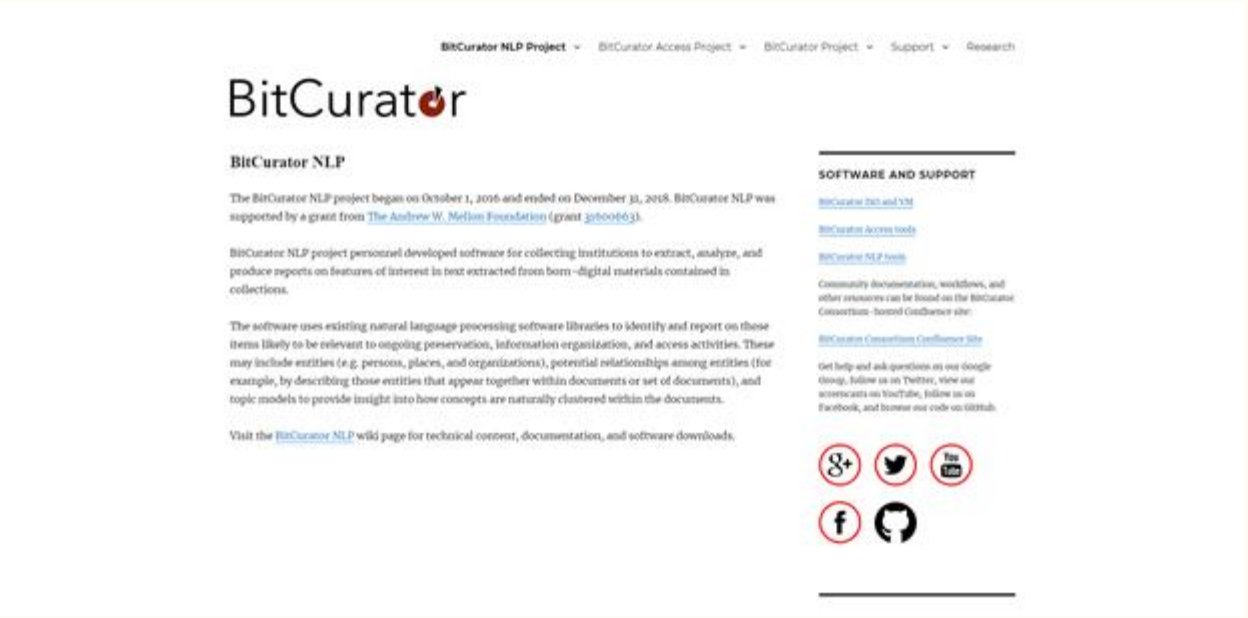
Další nástroje pro tvorbu a správu metadat

- METS Editor (SobekCM) – nástroj umožňuje vytváření METS záznamu např. ze složky obsahující soubory (https://sourceforge.net/projects/metseditor/). Umí vytvořit popisná metadata a strukturální mapy. Poslední verze je z roku 2015.

- Curator's Workbench – nástroj na vytváření METS záznamů, s vnořenými popisnými metadaty ve standardu MODS. Dokáže vytvořit METS záznam pro konkrétní objekty, s určenou strukturou apod. Nerozvíjen od 2013.

- PIMTOOLS – PREMIS in METS Toolbox – 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. Nerozvíjen.

Bitcurator (https://bitcurator.net/)

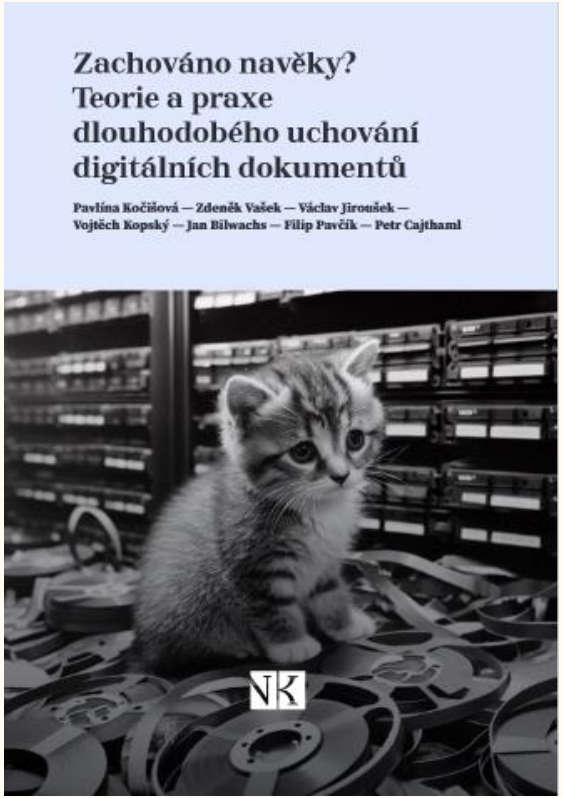


- Nástroj na tzv. digitální forenzní analýzu.
- Vytvoří:
- profily datasetů např. z pohledu formátů,
- přehledy metadat (last edited, created)
- provádí migrace některých formátů,
- vyhledává v datech údaje jako emaily, citlivé údaje apod.
- BitCurator je open source, je distribuován jako linux distribuce

LTP systémy v ČR v paměťových institucích

- LTP NDK
- ARCLib
- Národní digitální archiv (Archivematica)
- RODA (Archiv hl. města Prahy)
- Archiv UK (ve vývoji)

Zachováno navěky? Teorie a praxe dlouhodobého uchování digitálních dokumentů
Pavlína Kočišová – Zdeněk Vašek – Václav Jiroušek – Vojtěch Kopský – Jan Bilwachs – Filip Pavčík – Petr Cajthaml, Praha: Národní knihovna ČR, 2023, ISBN 978-80-7050-791-9.



DĚKUJI ZA POZORNOST!

zdenek.vasek@ruk.cuni.cz