LTP ochrana v Archivním informačním systému Univerzity Karlovy

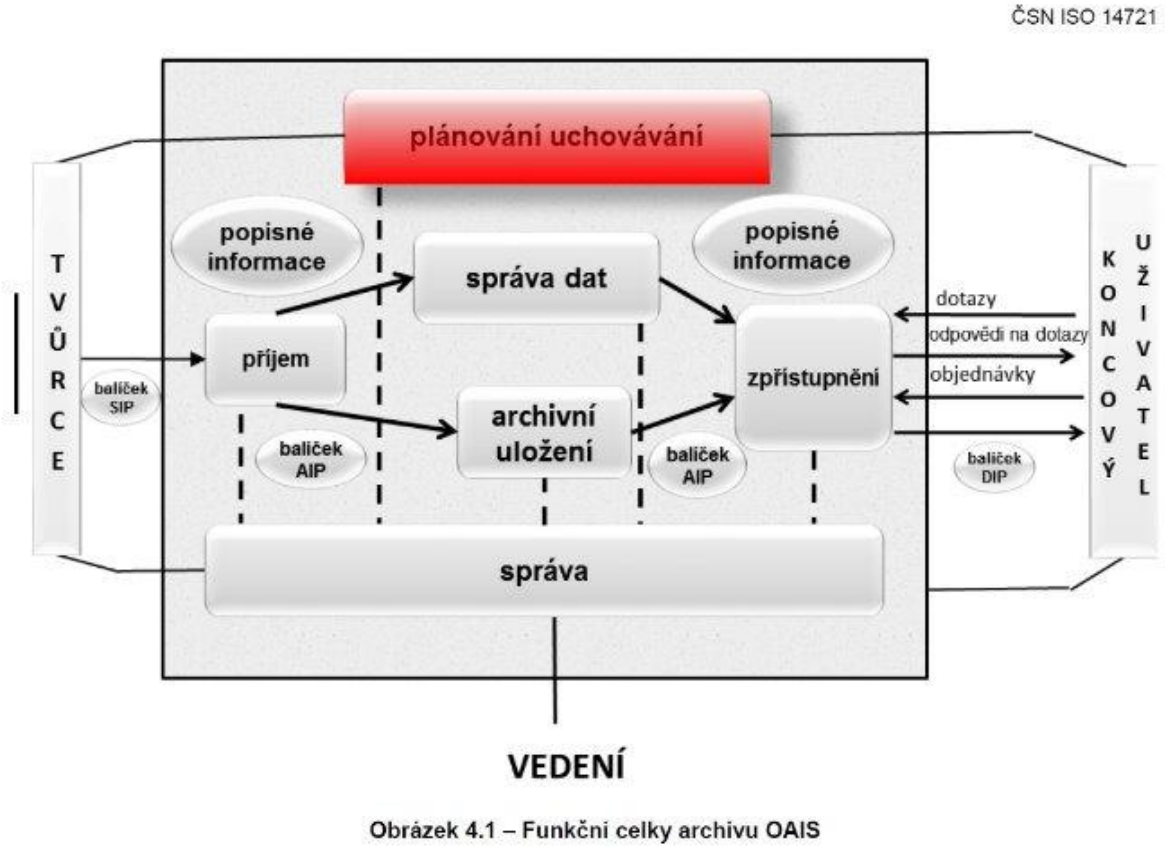
Petr Cajthaml a Zdeněk Vašek, Archiv Univerzity Karlovy

Univerzita Karlova

Uchování digitálních dokumentů v Archivu UK

Univerzita Karlova

- Archiv UK jako akreditovaný archiv
- Cílem Digitální archiv a trvalé uchování digitálních archiválií
- Součástí i utilita pro Long Term Preservation (OAIS)
- Hrozby: fyzické poškození, ztráta čitelnosti kvůli nemožnosti zobrazení a intepretace ⟶ bitová a logická vrstva ochrany



Obrázek 4.1 – Funkční celky archivu OAIS

Long Term Preservation - Dlouhodobé uchovávání

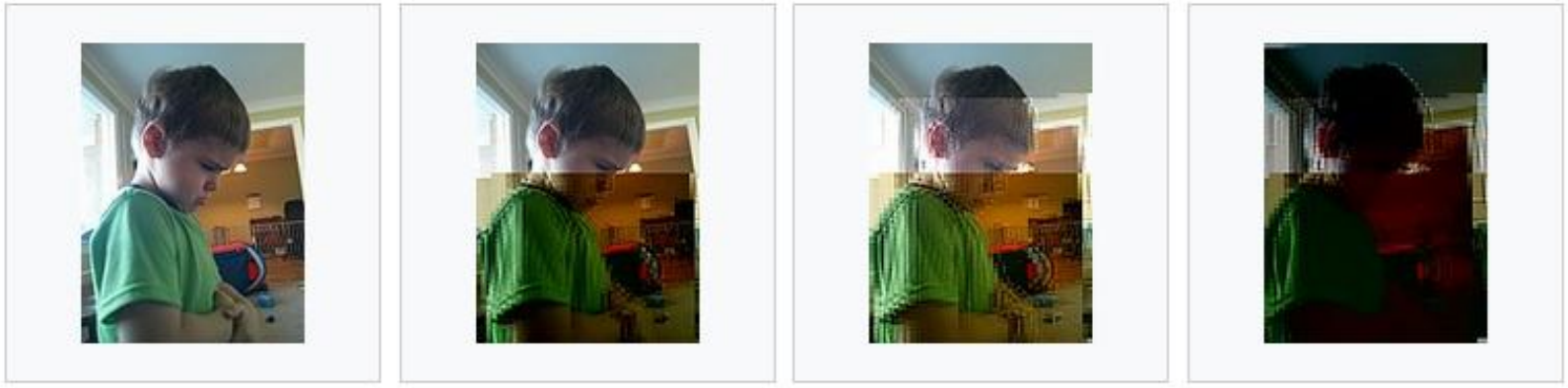
- **Dlouhodobé uchovávání (long term preservation):**

"dlouhodobé udržování informací – v podobě, která je určené skupině srozumitelná sama o sobě –, a dokladů o jejich hodnověrnosti" (ČSN ISO 14721 OAIS)

- Bitová ochrana

- Logická ochrana

"Abychom to otevřeli a …" (bitová ochrana)

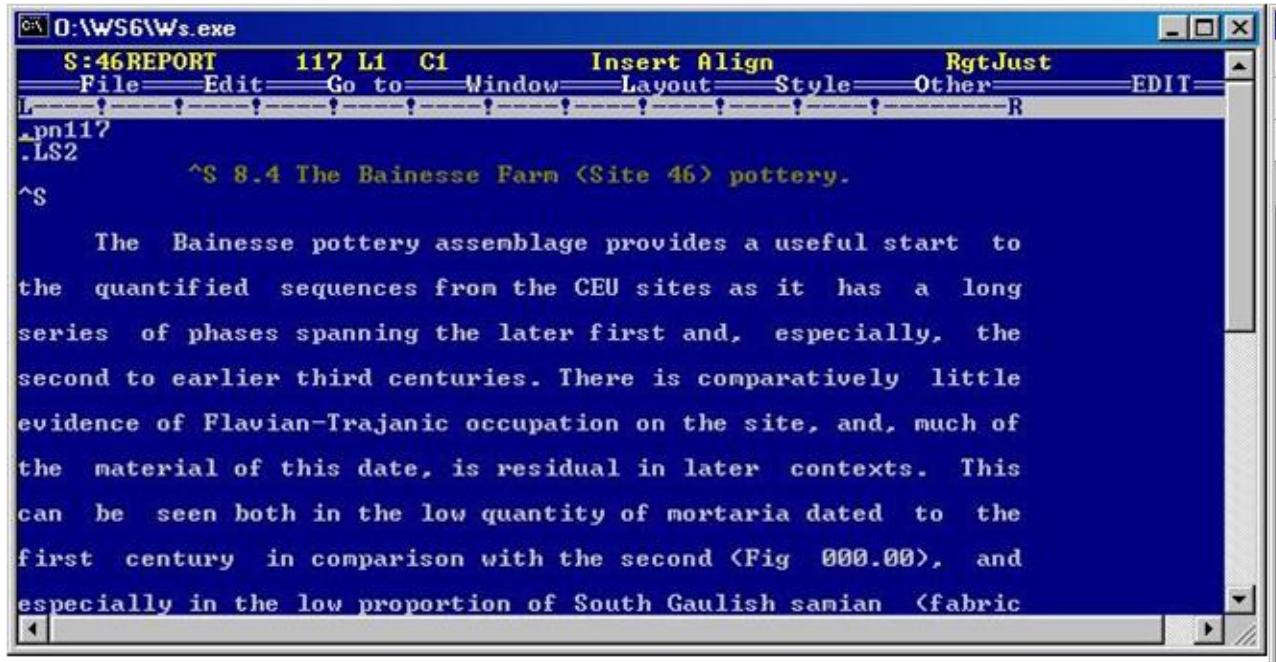


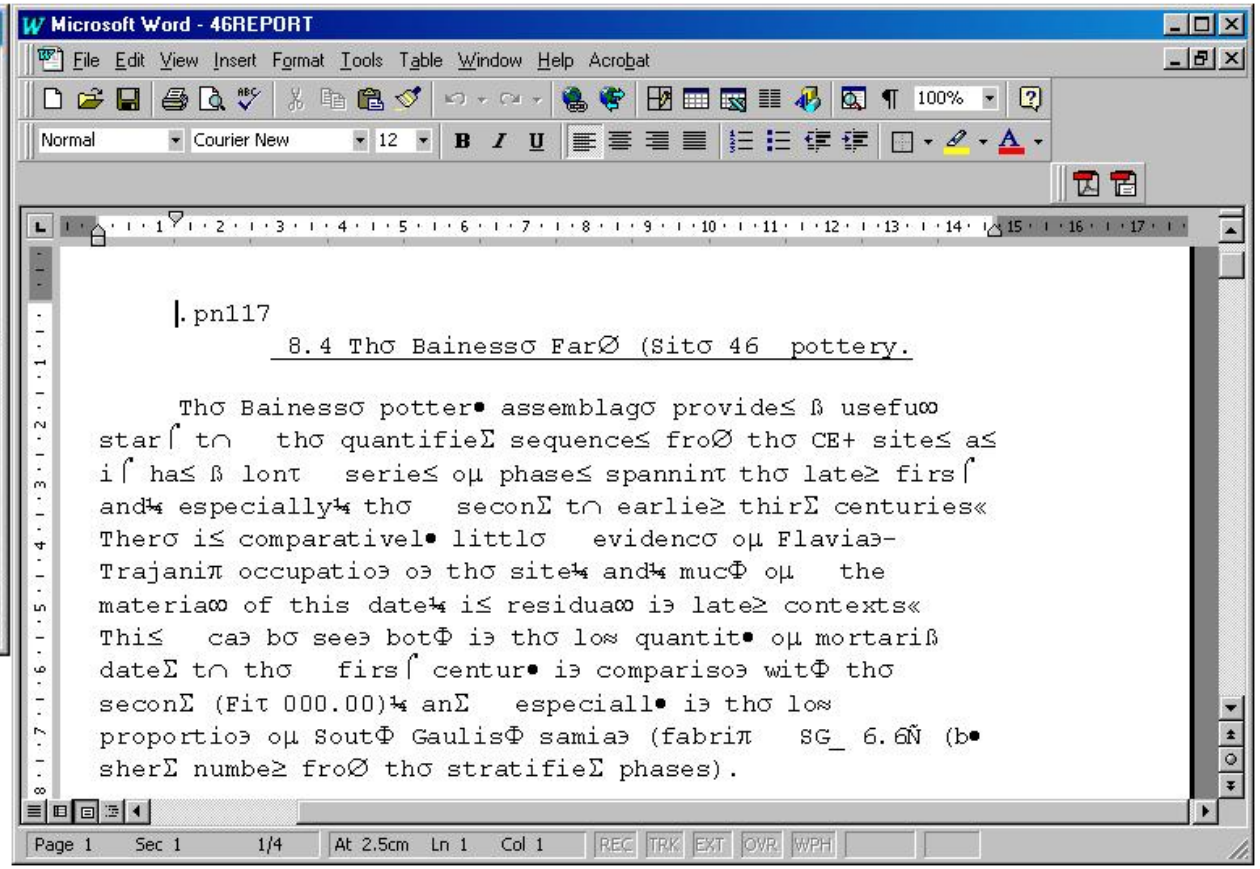
- Kontrolní součty
- Násobné kopie
- Více úložných technologií
- Vzdálené uložení

"… aby nám nezbyl rozsypaný čaj."

Univerzita Karlova

- Příklad Migrace WordStar do MS WORD (7 bit vs 8 bit)

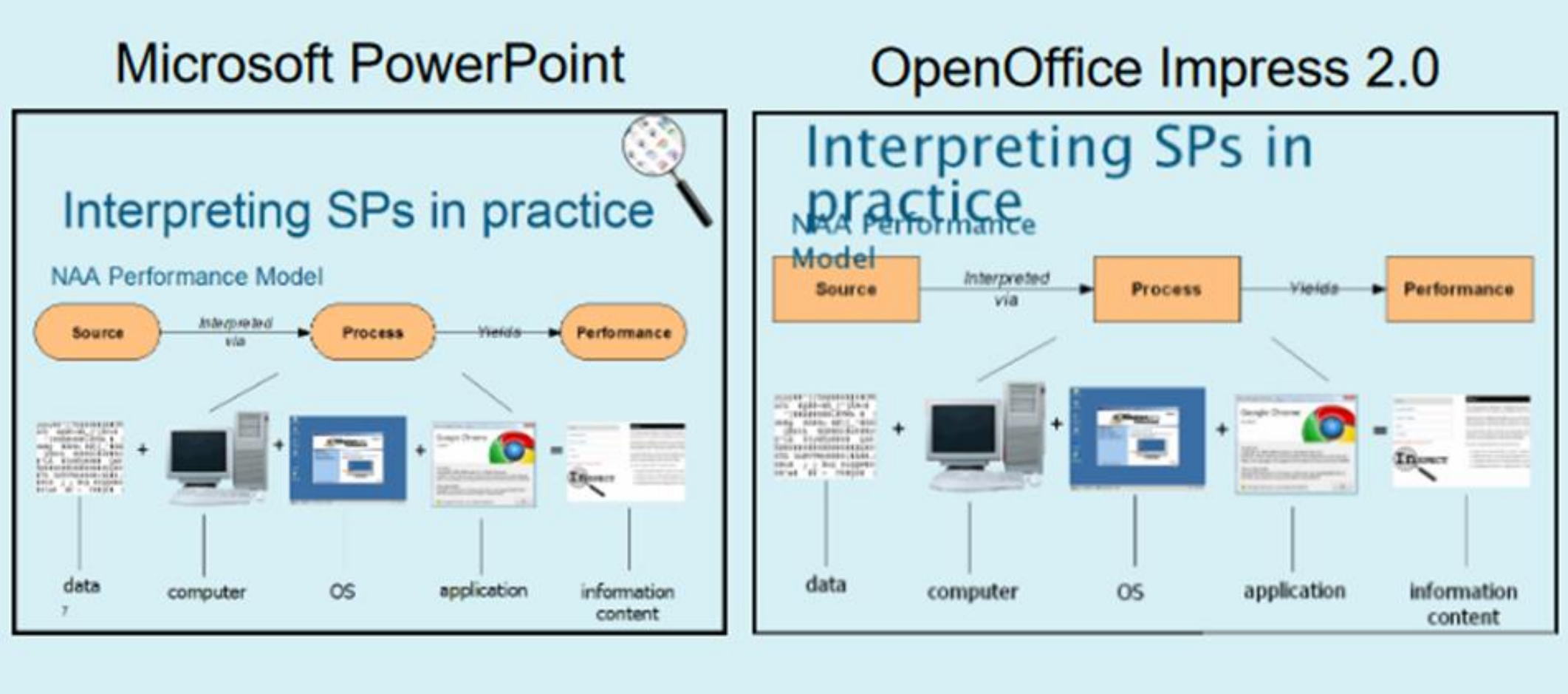




Logická ochrana - zachování informačního obsahu, určená komunita, signifikantní vlastnosti

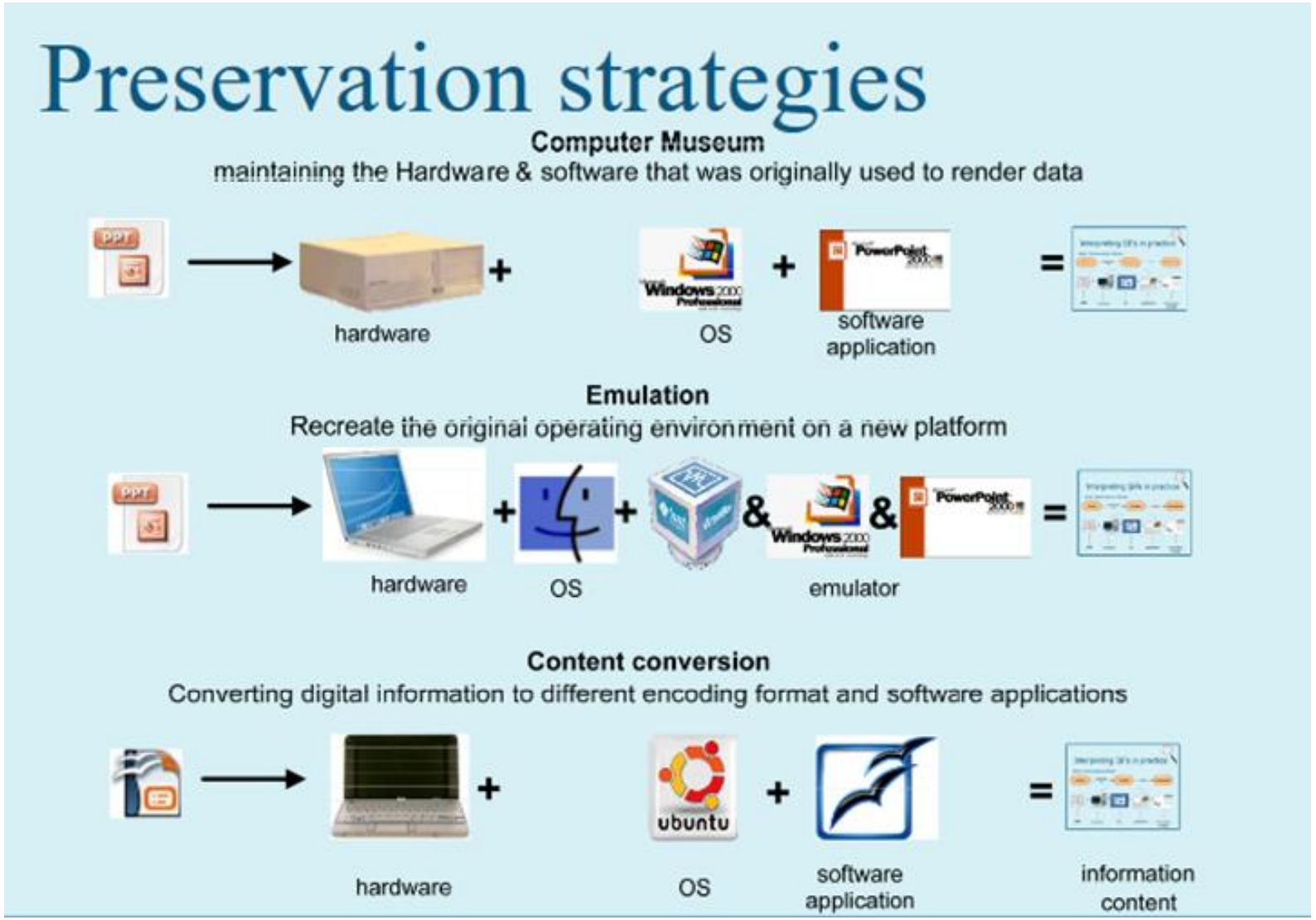
LTP = Uchování informačního obsahu i zachování formy

Univerzita Karlova



Strategie dlouhodobého uchovávání digitálních dokumentů

Univerzita Karlova



Logická ochrana v rámci předarchivní péče

Univerzita Karlova

- Nejméně nákladnou a nejúčinnější formou zajištění čitelnosti je standardizace dat a metadat již při vzniku nebo na vstupu do digitálního archivu
- Formátová omezení na výstupu u původce - legislativa (vyhláška 259/2012 Sb. o podrobnostech výkonu spisové služby) - výstupní datové formáty
- Interní předpisy UK, formátová pravidla Archivu UK - kvalifikační práce
- Nelze řešit vše u původce - občas se musí vzít cokoliv, nebo to zanikne

Realizace LTP v prostředí Archivu UK

Univerzita Karlova

- Modul LTP jako součást Archivního informačního systému
- Bitová ochrana
- Strategií formátové migrace, sledování zastarávání, ohrožení formátů
- Formátová knihovna
- Analýza
- Migrace
- Zapojené nástroje

Bitová ochrana

Univerzita Karlova

- Zrcadlená disková pole, RAID
- Dvě lokality (Praha a Plzeň)
- Od 2025 funkční úložiště v prostorách LF UK v Plzni
- Ukládání na offline média (LTO pásky)
- Správa úložišť pomocí ArcLib (ArchivalStorage)
- Mechanismy pro periodické kontroly dat na úložištích

Formátová knihovna – abychom věděli jak

Univerzita Karlova

- PRONOM
- FDD (LOC)
- Lokální sekce umožňující vlastní definici a podrobnější popis
- Slučování formátů, aktualizace registru
- Sekce pro správu dokumentace a plánování procesů, správa rizik

Plánování ve formátové knihovně

Univerzita Karlova

- Vlastní popis formátu
- Doporučené migrační nástroje
- Doporučené analytické nástroje
- Doporučené interpretační nástroje
- Stupeň rizika formátu
- Poznámka k riziku
- Popis ochrany
- Plánování operací
- Statistiky použití
- Odkazy na další související formáty

Jak prakticky realizujeme?

Univerzita Karlova

- Nástroj pro dávkovou práci s obsahem informačních balíčků (možnost zpracování samostatně i nad celým archivem)
- Samotný modul nic neanalyzuje ani nemigruje, jen připravuje balíčky a soubory, spouští nad nimi integrované nástroje a výsledek ukládá.
- Zapojení open source analytických a migračních nástrojů
- Otevřená architektura, zapojení jednotlivých nástrojů v kontejnerech typu podman
- Žádný nástroj není upraven, jen je zapojen
- Snadná údržba
- Možnost doplnění nových nástrojů

Formátová analýza - abychom věděli co

Univerzita Karlova

- Identifikace, charakterizace (vytěžení technických metadat), validace
- Siegfried
- VeraPDF
- ExifTool
- Jhove
- MediaInfo
- LXML

V detailu i celku

Univerzita Karlova

- Rozhodování nad jednotlivým souborem i dávkou na základě podrobných logů a vizualizace výsledků
- Možnost rozhodovat samostatně o každém i schválení jen částečného výstupu

Modifikované soubory

Identifikace	Přípona	Stav	Povolit odbavení	Workflow	YAML	XML	Log	F	V	T
a_00404867_dok00001_f002_v00001	tif	ODBAVENÝ	☑	1 - Základní formátová analýza	YAML	XML	Log	☑	☑	☑
a_00404867_dok00002_f002_v00001	tif	ODBAVENÝ	☑	1 - Základní formátová analýza	YAML	XML	Log	☑	☑	☑
a_00404867_dok00003_f002_v00001	tif	ODBAVENÝ	☑	1 - Základní formátová analýza	YAML	XML	Log	☑	☑	☑

Možnosti vizuálního porovnání

Univerzita Karlova

a_00401904_dok00001_f002_v00001 - YAML

Input

```
object:
  size: 423683
  checksum: f9515e9b8bd6d201a32e7b6cac4082b2b85183be28eb1
43045132515a66f2f71f733811005d325ed001c21bbc79044354e08d0
9051ff187e1fea759927dcf859
  checksumType: SHA512
  extension: .tif
format:
  identifiers:
- - type: AIS_FORMAT_MODULE
-   value: 617b0391-8156-4c41-848e-c31e93b012ee
- - type: MIME
-   value: image/tiff
  - type: PRONOM
    value: fmt/353
validation:
  valid: true
```

Output

```
object:
  size: 423683
  checksum: f9515e9b8bd6d201a32e7b6cac4082b2b85183be28eb1
43045132515a66f2f71f733811005d325ed001c21bbc79044354e08d0
9051ff187e1fea759927dcf859
  checksumType: SHA512
  extension: .tif
format:
  identifiers:
  - type: PRONOM
    value: fmt/353
+ - type: MIME
+   value: image/tiff
+ software:
+   name: Siegfried
+   version: 1.11.1
+ time: '2024-11-21T10:43:41.365803+00:00'
validation:
  valid: true
+ software:
+   name: Jhove
+   version: 1.28.0
+ time: '2024-11-21T10:43:42.421016+00:00'
+ tech:
+ software:
+   name: ExifTool
```

Co s tím pokud je to v krabičce?

Univerzita Karlova

- Kontejner je soubor se soubory
- Jak se dívat na obsah kontejneru?
- Vyndat vše nebo jen část obsahu
- Extrakce obsahu zipu – globe library
- Prázdný kontejner jako objekt v archivu

Co v archivu máme? Formát

Univerzita Karlova

Formát: Acrobat PDF/A - Portable Document Format 3a

Vytvořit dávku

Identifikátory	pdf, application/pdf, fmt/479
Další názvy	

PRONOM

Lokální sekce

Poslední statistiky

Identifikátor	6ae0a821-5330-48e0-b176-4d1b7e275128
Čas	1. 11. 2024 10:54:03

Nevnořené soubory

	Platné - počet	Všechny - počet	Platné - velikost	Všechny - velikost
Koncepty	0	0	0.00 KB	0.00 KB
Originály	9 591	9 591	9.07 GB	9.07 GB
LTP kopie	0	0	0.00 KB	0.00 KB
HQ kopie	0	0	0.00 KB	0.00 KB
LQ kopie	0	0	0.00 KB	0.00 KB
Metadata	0	0	0.00 KB	0.00 KB
Přílohy	22	25	3.70 MB	11.07 MB

Vnořené soubory

	Platné - počet	Všechny - počet	Platné - velikost	Všechny - velikost
Koncepty	0	0	0.00 KB	0.00 KB
Originály	302	302	260.12 MB	260.12 MB
LTP kopie	0	0	0.00 KB	0.00 KB
HQ kopie	0	0	0.00 KB	0.00 KB
LQ kopie	0	0	0.00 KB	0.00 KB
Metadata	0	0	0.00 KB	0.00 KB
Přílohy	0	0	0.00 KB	0.00 KB

Co v archivu máme? Archivní soubor a jeho data

Univerzita Karlova

Název	Stupeň rizika	Originály - počet
Acrobat PDF/A - Portable Document Format 1a	1	2 566
Acrobat PDF 1.5 - Portable Document Format 1.5	1	2 096
Acrobat PDF 1.7 - Portable Document Format 1.7	1	1 457
Extensible Markup Language 1.0	1	1 428
Acrobat PDF 1.4 - Portable Document Format 1.4	1	671
Acrobat PDF/A - Portable Document Format 3a	1	629
Acrobat PDF/A - Portable Document Format 1b	1	480
Acrobat PDF 1.6 - Portable Document Format 1.6	1	369
Acrobat PDF 1.3 - Portable Document Format 1.3	1	252
Acrobat PDF/A - Portable Document Format 2b	1	145
Acrobat PDF/A - Portable Document Format 2u	1	30
Acrobat PDF/A - Portable Document Format 3b	1	24
ZIP Format	1	10
JPEG File Interchange Format 1.01	1	9
RAR Archive 2.9	1	3
Microsoft Word for Windows 2007 onwards	1	3
Exchangeable Image File Format (Compressed) 2.3.x	1	2
JPEG File Interchange Format 1.02	1	2
Microsoft Excel for Windows 2007 onwards	1	1
Plain Text File	1	1
Exchangeable Image File Format (Compressed) 2.21	1	1
Microsoft Excel 97 Workbook (xls) 8	1	0
Comma Separated Values	1	0

Co s tím? Zmigrovat!

Univerzita Karlova

- Aktuálně 314 formátů s alespoň jedním souborem
- Připravená workflow
- Integrované nástroje, možnost rozšíření
- Možnosti uživatelského nastavení a kontroly, podrobné logy
- Součástí workflow vždy ověření výsledku

Čím migrujeme?

Univerzita Karlova

- PDF - ghostsript
- obrazové formáty - GraphicsMagick
- audio/video formáty - FFmpeg
- textové dokumenty - LibreOffice
- tabulkové soubory - LibreOffice Calc

Čísla a plány

Univerzita Karlova

- Dodavatel AipSafe
- Vývoj 05/2023 - 05/2024
- Testování 05 – 09/2024
- Nasazeno od 10/2024
- Cena za vývoj 2,4 mil. Kč vč. DPH
- 2025 - dokončení dokumentace a žádost o akreditaci Dig. archivu

Dokumentace - otevřený systém

https://frnk.lightcomp.cz/download/cuni-ais/doc/fm/fm.html

Univerzita Karlova

Search docs

OBSAH

1. Novinky v dokumentaci
2. Administrace
3. Specifikace datových formátů a sdílených služeb
4. Modul Výběr
5. Modul UPL
6. Evidenčně Správní Modul
7. Formátový modul
   - 7.1. Administrace
   - 7.2. Integrace
   - 7.3. Testování
   - 7.4. Uživatelská příručka
     - 7.4.1. Uživatelská oprávnění
     - 7.4.2. Knihovna
     - 7.4.3. Formátová analýza
       - Práce s workflow
       - Práce s procesy formátové analýzy
       - Práce s balíčky
       - Práce se soubory
       - Práce ve workspace
     - 7.4.4. Formátová migrace
8. Modul pro offline média
9. Elza
10. Imwhooser
11. Modul Zápis
12. Archival Storage
13. Badatelna
14. Technická dokumentace
15. Tiskové šablony

Struktura pracovního prostoru

Pracovní prostor obsahuje adrsáře pro jednotlivé procesy. Jméno adresáře procesu je dáno identifikací procesu.

Uvnitř adresáře procesu jsou adresáře zpracovávaných balíčků. Jméno adresáře balíčku je dáno jeho id na úložišti.

Adresář balíčku se skládá z následujících podadresářů:

- input-files - Obsahuje vstupní soubory do formátové analýzy,
- input-metadata - Obsahuje původní metadata souborů.
- output-metadata - Obsahuje nová metadata souborů a log jejich zpracování.

V adresářích je každý soubor pojmenován svým AIS ID (a_00...) s příslušnou příponou.

Metadata jsou uvedena ve dvou souborech yml a xml. YML soubor má pevnou strukturu stejnou pro všechny typy souborů. XML soubor obsahuje obecná technická metadata závislá na formátu souboru umísťovaná do premis objectCharacteristicsExtension.

Struktura YML souboru

```yaml
format: # Výstup identifikace formátu
  identifiers: # seznam - Seznam identifikátorů rozpoznaného formátu
    - type: PRONOM # string Typ identifikátoru (PRONOM, FDD, CUSTOM)
      value: fmt/17 # string Hodnota identifikátoru
    - type: MIME
      value: application/pdf
  software:
    name: Siegfried # string - Jméno nástroje, který identifikaci formátu provedl.
    version: 1.10.1 # string - Verze nástroje, který identifikaci formátu provedl.
  time: '2023-11-10T13:26:20.235020+01:00' # string - Datum a čas identifikace formátu v iso 8601 formátu.
validation: # Výstup validace
  valid: false # boolean - Výstup validace.
  messages: # seznam stringů - Zprávy upřesňující výstup validace.
    - The value of Author entry from the document Info dictionary and its matching XMP
      property dc:creator are not equivalent (Info /Author = Philip Hutchison, XMP dc:creator
      = null)
    - The document catalog dictionary doesn't contain metadata key.
    - The value of Creator entry from the document Info dictionary and its matching
      XMP property xmp:CreatorTool are not equivalent (Info /Creator = Pages, XMP xmp:CreatorTool
      = null)
    - The value of Producer entry from the document Info dictionary and its matching
      XMP property pdf:Producer are not equivalent (Info /Producer = Mac OS X 10.5.4
      Quartz PDFContext, XMP pdf:Producer = null)
    - The value of Title entry from the document Info dictionary and its matching XMP
      property dc:title['x-default'] are not equivalent (Info /Title = sample, XMP dc:title['x-default']
      = null)
  software:
    name: VeraPDF # string - Jméno nástroje, který validaci provedl.
    version: 1.24.1 # string - Verze nástroje, který validaci provedl.
  time: '2023-11-10T13:26:22.835225+01:00' # string - Datum a čas validace v iso 8601 formátu.
tech: # Výstup extrakce technických metadat, samotná metadata jsou v xml souboru.
```

Děkujeme za pozornost!

https://ais.udauk.cuni.cz/

https://frnk.lightcomp.cz/download/cuni-ais/doc/fm/fm.html

Petr Cajthaml
petr.cajthaml@ruk.cuni.cz

Zdeněk Vašek
zdenek.vasek@ruk.cuni.cz

Univerzita Karlova