Digitalia

Digital preservation ”workflow” with open-source tools

Ph.D Anssi Jääskeläinen

Research manager Xamk/Digitalia

Anssi.jaaskelainen@xamk.fi

XAMK

South-Eastern Finland University of Applied Sciences

XAMK

Memory Lab

Agenda

- Xamk, Digitalia, Memory Lab.
- Commercial solutions
- Open-source ways & licensing
- **Toolset for the workflow**:
  - Worst case scenario: Unidentified files that needs to be preserved in a proper format inside a proper IP package

Xamk / Digitalia / Memory Lab

- South-Eastern Finland University of Applied Sciences
  - https://www.xamk.fi/en/frontpage/
- Memory Lab
  - AI specialized ~680 000€ technical environment
  - Operational June 23.
- Digitalia – Research Center on Digital Information Management
  - Usability of digital materials
  - Automated things
  - Visualization
  - Digitalia.fi
  - More things: https://digitalia.xamk.fi/

Commercial / Proprietary solutions

- Everything ”just” works
- Generally trustworthy
- In most cases the only way
- Challenges
  - Price
  - Limited features
  - Export
  - Integrations

Open-source way

- “Do one thing and do it well”
- Requires inhouse knowhow
- Stackoverflow and other discussion communities are your best friends
- Simple to swap components
- Code is freely available for modifications
  - Licensing!

Workflow

1. Identifying filetypes
2. Finding the correct preservation format
3. Migration to preservation format
4. Validating the preservation format
5. Creating a SIP package
6. Ingesting SIP into an eark conformant archive

Identifying filetypes

Droid

- Built with Java
- https://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/
  - https://github.com/digital-preservation/droid
- Identifies “a lot of” file types
- Linked with PRONOM
- Exports csv

Droid command

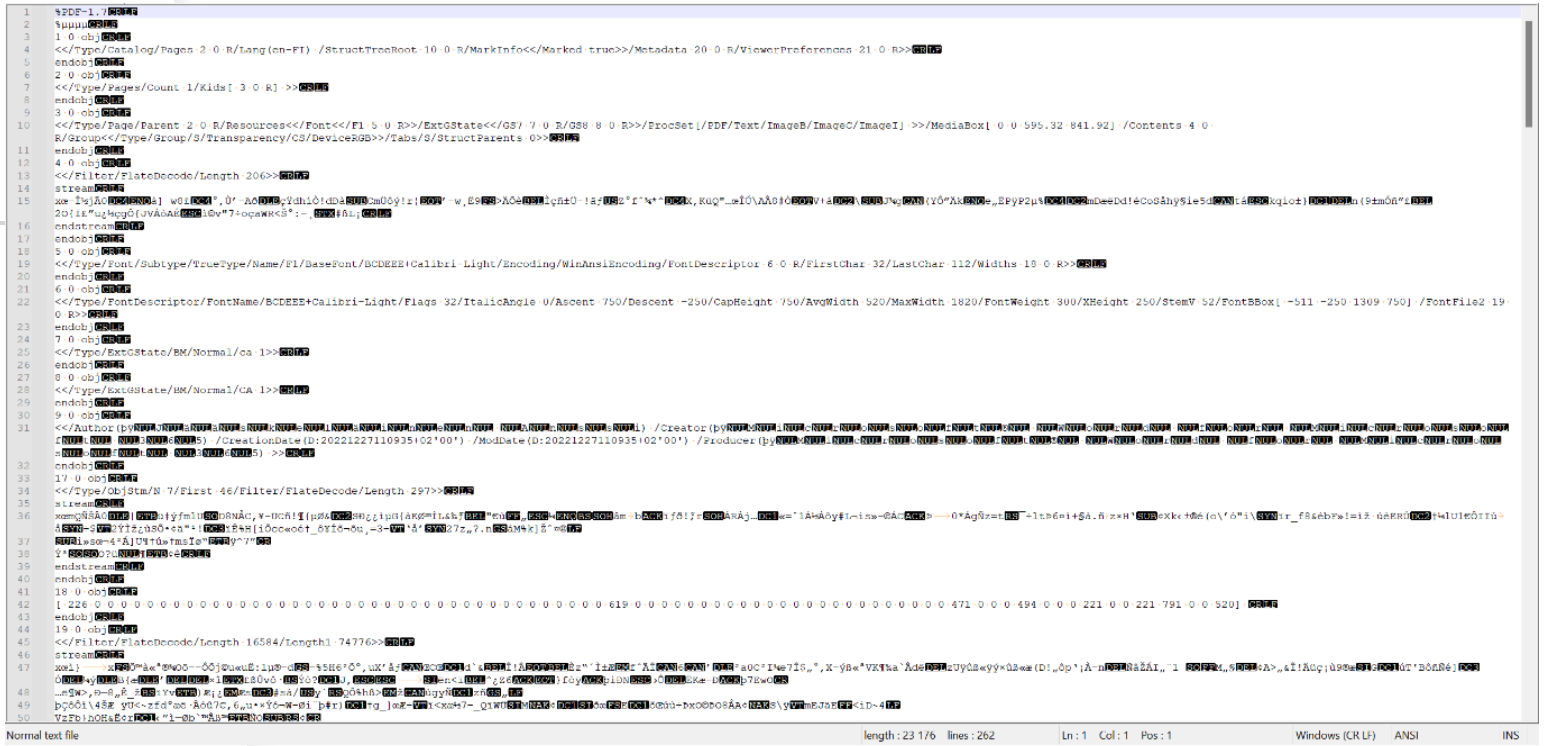
- droidcmd = [droid_path, '-Nr', fullFilePath, '-Ns', droid_signature_path, '-Nc', droid_container_path, '--quiet’]
- droidprocess= subprocess.Popen(droidcmd, stdout=subprocess.PIPE, universal_newlines=True)

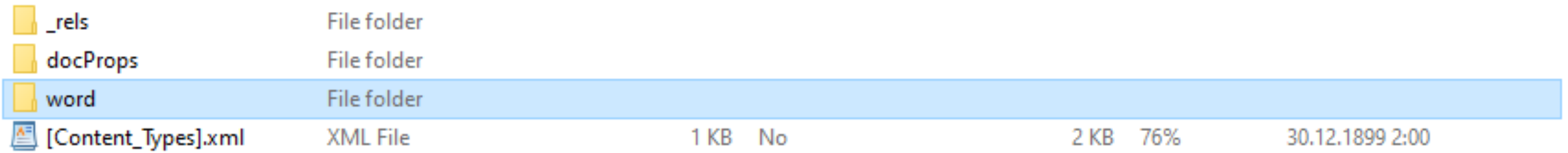
Droid demo

5 unidentified files

What to do when Droid fails

- Apache Tika to see if metadata reveals something
- Notepad++ (or other competent) text editor to see the actual content
- Unzip, etc. compression tool



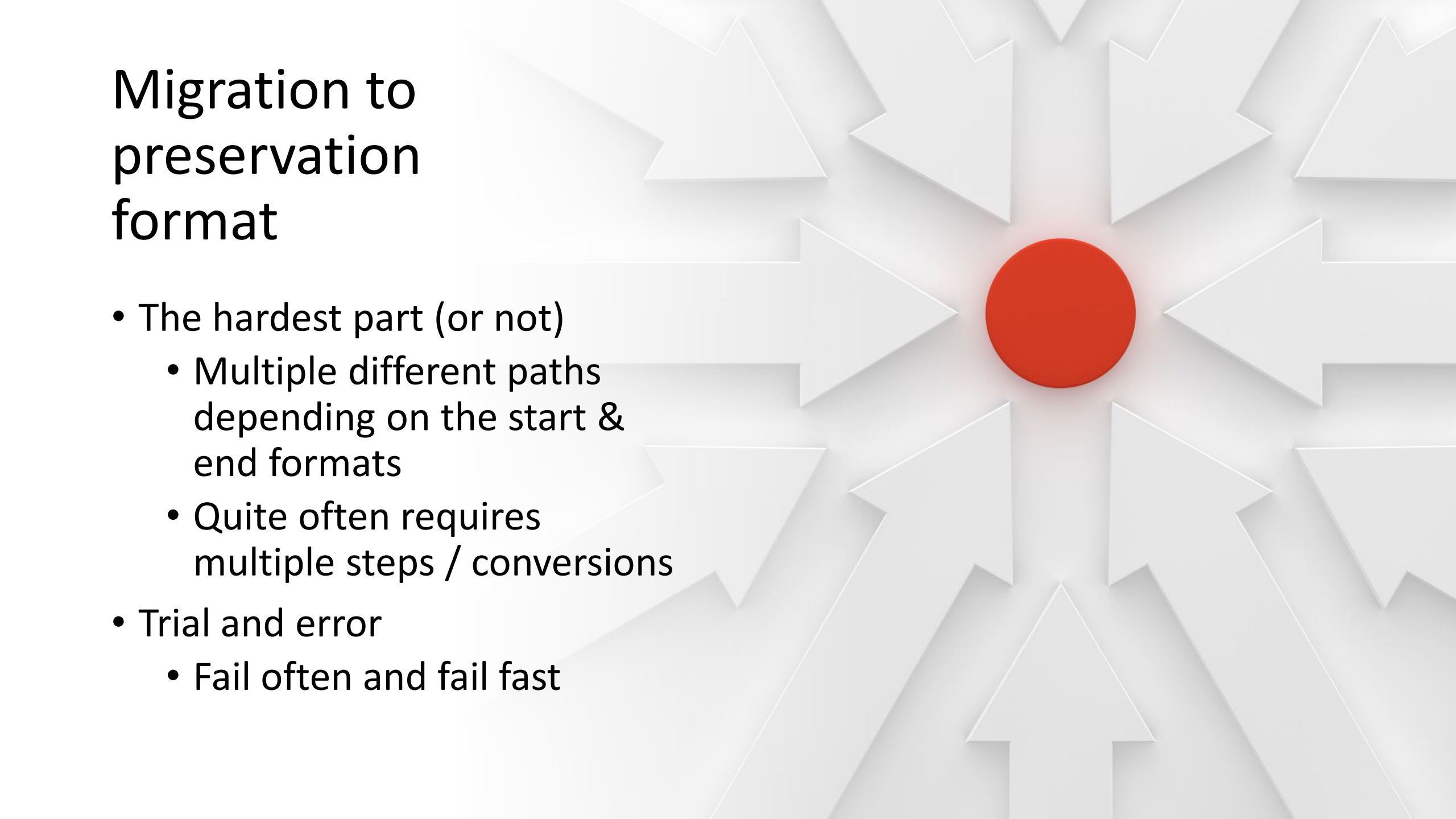


Finding “the right” preservation format

- Manual step
- Many lists of preferred and accepted formats
  - LoC: https://www.loc.gov/preservation/resources/rfs/
  - Nara: https://www.archives.gov/records-mgmt/policy/transfer-guidance-tables.html
  - CSC: https://digitalpreservation.fi/en/specifications/fileformats
- Use generally accepted widely used file formats
  - It might be good idea to preserve also the editable format

Migration to preservation format

- The hardest part (or not)
  - Multiple different paths depending on the start & end formats
  - Quite often requires multiple steps / conversions
- Trial and error
  - Fail often and fail fast



IDENTIFIED FORMAT	TOOL	PRESERVATION FORMAT
Pdf (text)	Libreoffice draw / Ghostscript	pdf/a-3b
Bmp (image)	Gimp / ImageMagick	png
Mdb (database)	DBPTK	siard
Doc (text)	Libreoffice writer /MS Word / Abiword	pdf/a-3b
Ra (audio)	Ffmpeg / VLC-media player / Audacity	WAV

Conversion demo

- 5 identified files to preservation formats
- All can be accomplished via command line

```
cmd_gs = [ghostscript_path, '-dPDFA=3',
      '-dBATCH', '-dNOPAUSE', '-dNOOUTERSAVE', '-dNOSAFER', '-dPDFSETTINGS=/prepress',
      '-dPDFACompatibilityPolicy=1', '-dAutoFilterColorImages=false', '-dColorImageFilter=/FlateEncode',
      '-dAutoFilterGrayImages=false', '-dGrayImageFilter=/FlateEncode', '-dMonoImageFilter=/FlateEncode',
      '-sColorConversionStrategy=UseDeviceIndependentColor','-dEmbedAllFonts=true',
      '-sDEVICE=pdfwrite', outputF, 'pdfa_def.ps', pdf_file]
```

Validating the preservation format

- First step is to find the validator
- Validators
  - Verapdf
  - Jhove
    - Gif, html, jp2, pdf, png, tiff, wav, xml, etc.
  - Kost-Val
  - DBPTK
  - GitHub & Google

Creating a SIP package

Roda-In

ESSArch

Earkweb?

There is a simple alternative

- OneClick eArchiving
- SIP creator
  - Demo: https://digitalia.xamk.fi/oneclickUploader/uploader-main.php
  - Codes & tutorials: https://github.com/xamkfi/Digitalia-oneclick-full

SIP creator demo & upload

- Create a SIP package
- Upload the package to Roda
  - https://www.roda-community.org/#welcome
  - admin/roda

Parts that were missing from the “workflow”

- Virus checks
- Possible metadata conversions
- SIP package validation (OneClick uses CommonsIP https://github.com/keeps/commons-ip for validation)

List of apps

- Droid: https://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/
- Tika: https://tika.apache.org/
- Notepad++: https://notepad-plus-plus.org/
- LibreOffice: https://www.libreoffice.org/
- Gimp: https://www.gimp.org/
- Ghostscript: https://www.ghostscript.com/
- Jhove: https://jhove.openpreservation.org/
- DBPTK: https://database-preservation.com/
- Roda In: https://rodain.roda-community.org/
- ImageMagick: https://imagemagick.org/
- VeraPDF: https://verapdf.org/
- Ffmpeg: https://ffmpeg.org/
- VLC media player: https://www.videolan.org/vlc/
- Audacity: https://www.audacityteam.org/
- Abiword: https://www.abisource.com/
- KOST-Val: https://coptr.digipres.org/index.php/KOST-Val
- ClamAV: https://www.clamav.net/
- CommonsIP: https://github.com/keeps/commons-ip
- Roda: https://www.roda-community.org/#welcome


Questions, comments, criticism, worries, etc.?