Welcome to this live webinar on

Using open-source tools for eArchiving

Start 10:00

21 September 2023

Audience notes for the Live Webinar

Your **cameras have been turned off** and **microphones muted**.

If you have any technical issues during the event, please use the chat function.

Please **use the Q&A for questions to speakers**. These will be addressed at the end of the event.

Please note that this webinar is recorded. No attendee personal information will be captured in these recordings.

Agenda

10:00 – 10:05

eArchiving Initiative welcome

Jaime Kaminski – eArchiving Initiative training activity lead

10:05 – 10:50

Using open source tools for eArchiving

Anssi Jääskeläinen – South-Eastern Finland University of Applied Sciences, XAMK

10:50 – 11:00

Q&A

Using open source tools for eArchiving

Anssi Jääskeläinen, South-Eastern Finland University of Applied Sciences, XAMK

eArchiving Initiative Training Webinar

European Commission

Using open-source tools for eArchiving

Ph.D Anssi Jääskeläinen

Research manager

Anssi.jaaskelainen@xamk.fi

Xamk/Digitalia

Digitalia

XAMK

XAMK

Memory Lab

Agenda

- Xamk, Digitalia, Memory Lab.
- Proprietary solutions
- Open source as an alternative
- Worst case scenario

European Commission

Xamk / Digitalia / Memory Lab

- South-Eastern Finland University of Applied Sciences
  - https://www.xamk.fi/en/frontpage/

- Memory Lab
  - AI specialised ~680 000€ technical environment
  - Fully installed and operational
  - Memorylab.fi published later on

- Digitalia – Research Center on Digital Information Management
  - Usability of digital materials
  - Automated things
  - Visualisation
  - Digitalia.fi
  - More things: https://digitalia.xamk.fi/

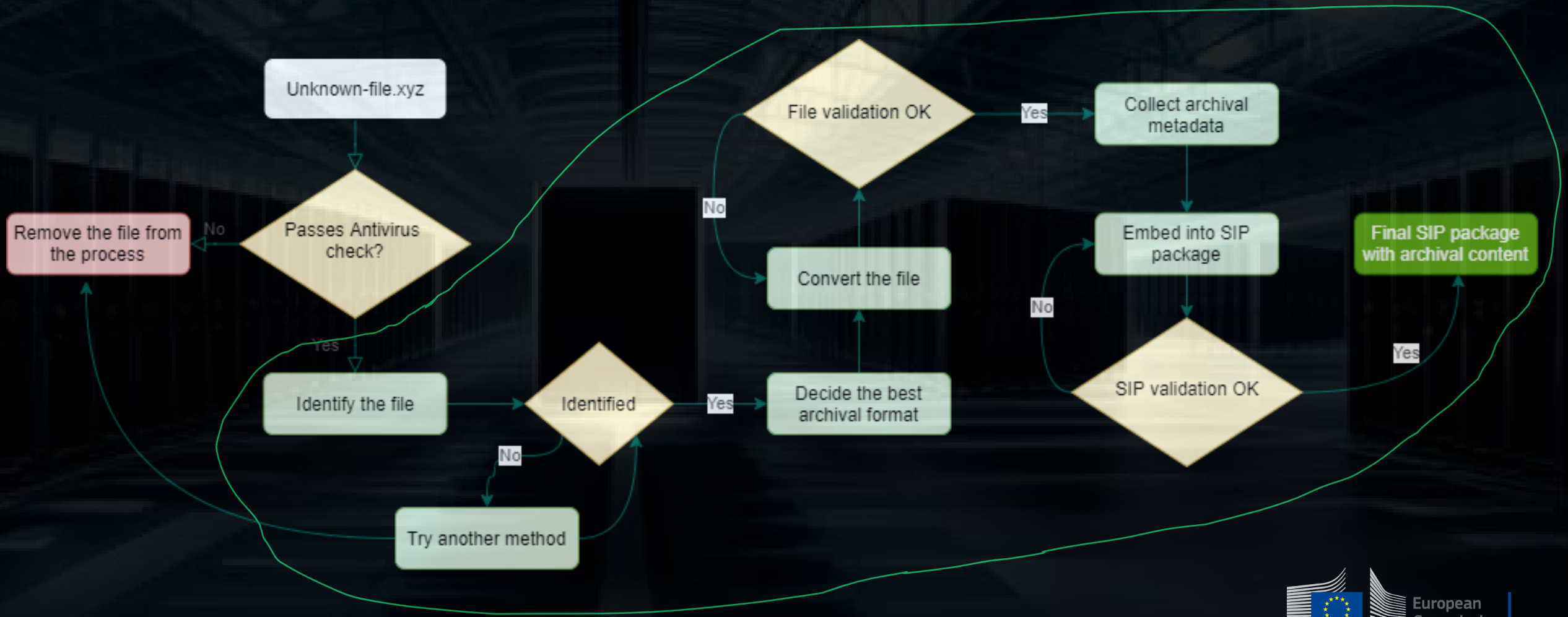
Commercial / Proprietary solutions

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

Open Source != silver bullet

- “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

Worst case workflow



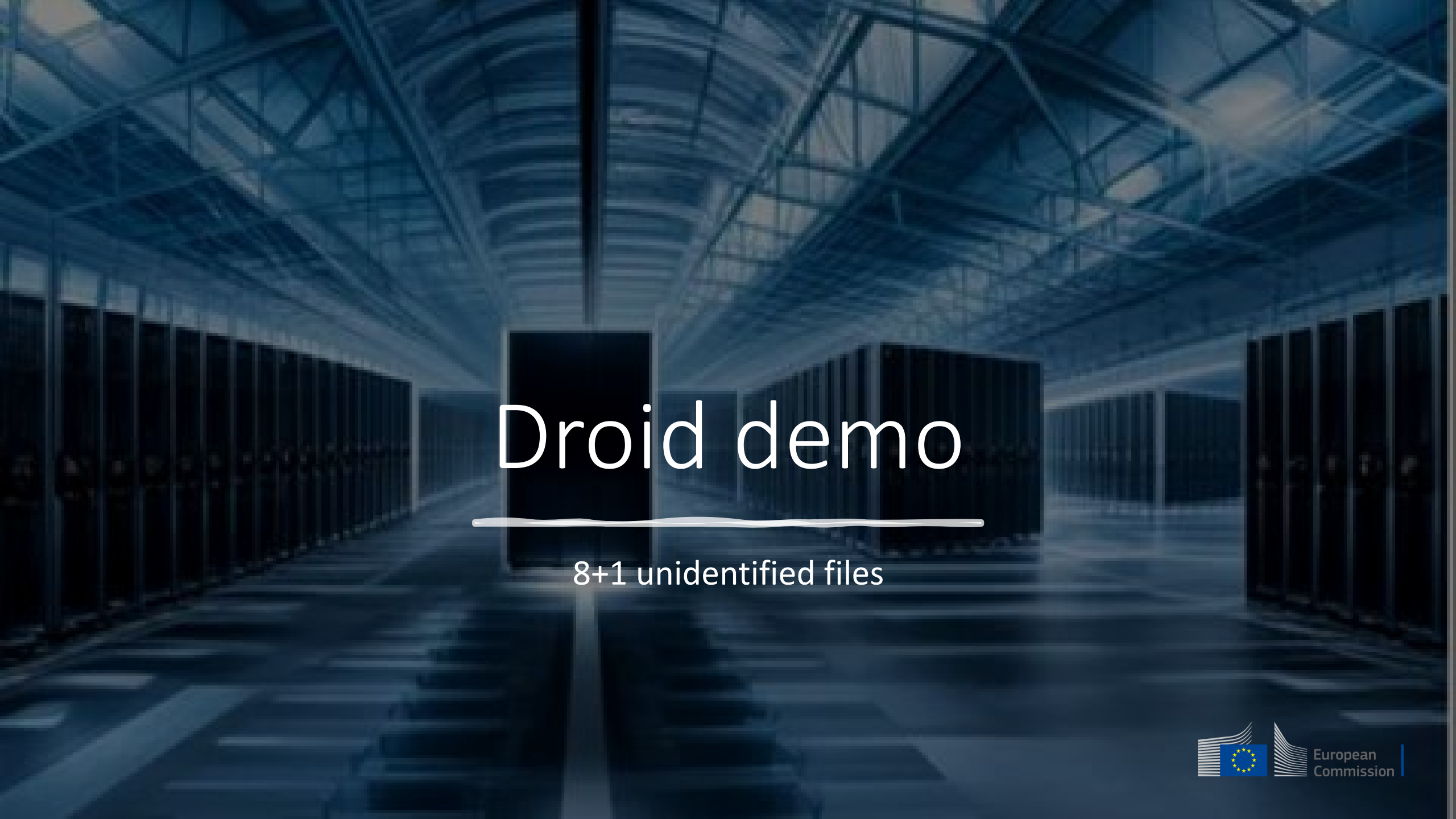
European Commission

Tools for building OS workflow

- dbptk
- droid
- GIMP
- HandBrake
- ImageMagick
- jhove
- KOST-Val
- LibreOffice 7.4
- Notepad++
- RODA-in
- tika
- verapdf
- virus scanner

Identify the file

- Using Droid (The National Archives, UK) https://github.com/digital-preservation/droid
  - Implemented with Java
  - Linked to Pronom database, https://www.nationalarchives.gov.uk/PRONOM/
- UI & CLI
- java -jar droid-command-line-6.6.1.jar sample2
  - "ID","PARENT_ID","URI","FILE_PATH","NAME","METHOD","STATUS","SIZE","TYPE","EXT","LAST_MODIFIED","EXTENSION_MISMATCH","HASH","FORMAT_COUNT","PUID","MIME_TYPE","FORMAT_NAME","FORMAT_VERSION"
  - "1","","file:/home/digitalia-aj/Downloads/sample2","/home/digitalia-aj/Downloads/sample2","sample2","Signature","Done","4129","File","","2023-09-15T11:38:42","true","","1","fmt/396","","PocketMobi (Palm Resource) File",""



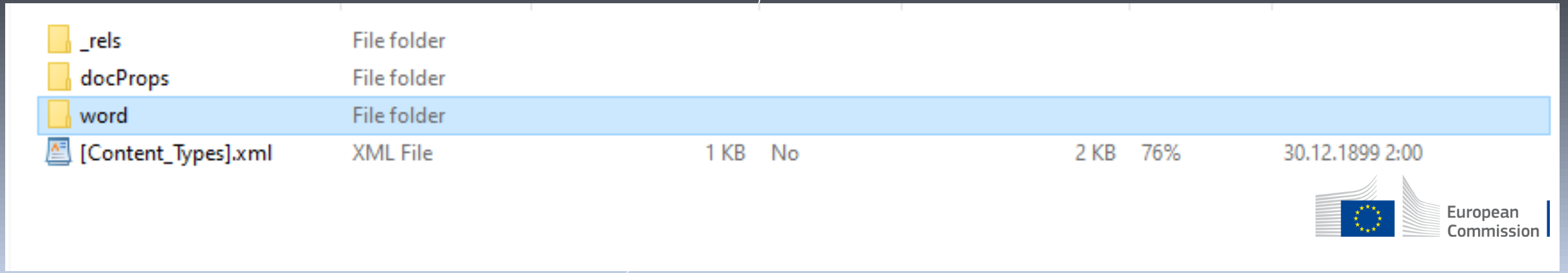
Droid demo

8+1 unidentified files

European Commission

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



European Commission

Decide the best archival format

- Manual step
- Follow 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
  - OPF International Comparison of Recommended File Formats - Google Sheets
- Use generally accepted widely used file formats
  - It might be good idea to preserve also the editable format

Archival formats and tools

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

Convert the file

- Probably the hardest part
  - Multiple different paths depending on the start and end formats
  - Sometimes requires multiple steps / conversions
- Trial and error
  - Fail often and fail fast

Conversion demo

- files to preservation format
- 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]
```

File validation

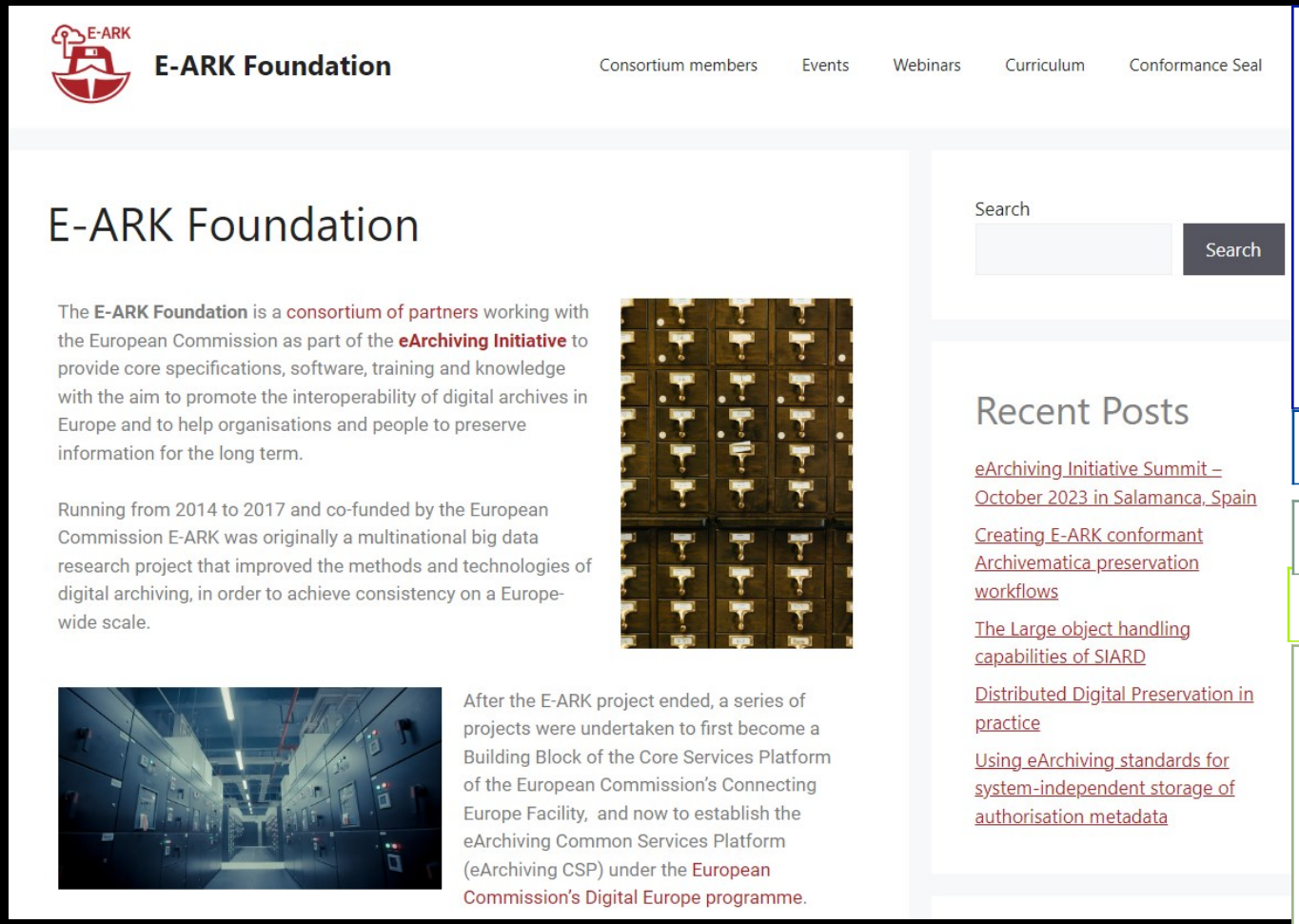
- First step is to find a validator
- Validators
  - Verapdf
  - Jhove
    - Gif, html, jp2, pdf, png, tiff, wav, xml, etc.
  - Kost-Val
  - DBPTK
  - GitHub and Google

European Commission

Collect archival metadata

- No single tool but
  - Tika / Exiftool for finding technical metadata
- https://arkkiivi.fi/ for content related metadata
  - Works best with Finnish but has Swedish and English support
  - Has some nice additional features

Collect archival metadata



Organisations	GEARK E-ARK Foundation, the European Commission, the eArchiving Ini, -, Digital Europe, the European Commission's, E, The E-ARK Foundation, E-ARK Foundation, ARK
Geopolitical locations	Salamanca, Spain
Date	2014 to 2017, October 2023
Locations	Europe
Index terms	digital technology, electronic archiving, data storage, webinars, software technology, research programmes, projects, postings (work orders), internship, electronic archives

Creating a SIP package

Roda-In – https://github.com/keeps/roda-in

ESSArch – https://github.com/ESSolutions/ESSArch

Earkweb – https://github.com/E-ARK-Software/earkweb

The most simple option

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

European Commission

SIP creator demo and upload

Create a SIP package

SIP to Roda

- https://www.roda-community.org/#welcome
- admin/roda

European Commission

Parts that were missing from the “workflow”

- Virus checks
- Metadata type conversions
- SIP package validation (OneClick uses CommonsIP https://github.com/keeps/commons-ip for validation)
- All nice technical details, programming, malfunctioning apps, broken libraries etc.

European Commission

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.?

Contact:
Anssi.Jaaskelainen@xamk.fi
Linkedin
Digitalia.fi

European Commission

Thank you

Contact

https://e-ark4all.eu/

info@e-ark-foundation.com

@EU_eArchiving

https://www.linkedin.com/company/eu-earchiving-initiative

https://www.youtube.com/@e-ark

Thank you

European Commission