The challenge of
Digital Preservation at CERN

Antonio Vivace // PV2023 - May 3rd 2023 // avivace@cern.ch

Contents

1. Preservation Scope
2. Scenarios
3. Digital Preservation
4. Creating SIPs
5. Platform
    a. Architecture overview
    b. Features
    c. Technology
6. Further improvements

Preservation Scope

- Digital Repositories in use at CERN
- Local folders (user provided content)
    - E.g. Slides submitted to external conferences, notes, drafts

GitLab

zenodo

opendata

CDS
CERN Document Server

CodiMD

indico

NOT

Another digital repository

A backup

But...

Policies, **infrastructures** and **technologies** to face challenges of file corruption, media failure and technological (hardware and software) obsolescence, following OAIS principles

Scenarios

A. Repositories periodically selecting and submitting resources for long term preservation
   - service implements preservation (AIPS) and register them to DM platform
   - service **submits** SIPs to DM platform
   - service request DM platform to **harvest** their resources

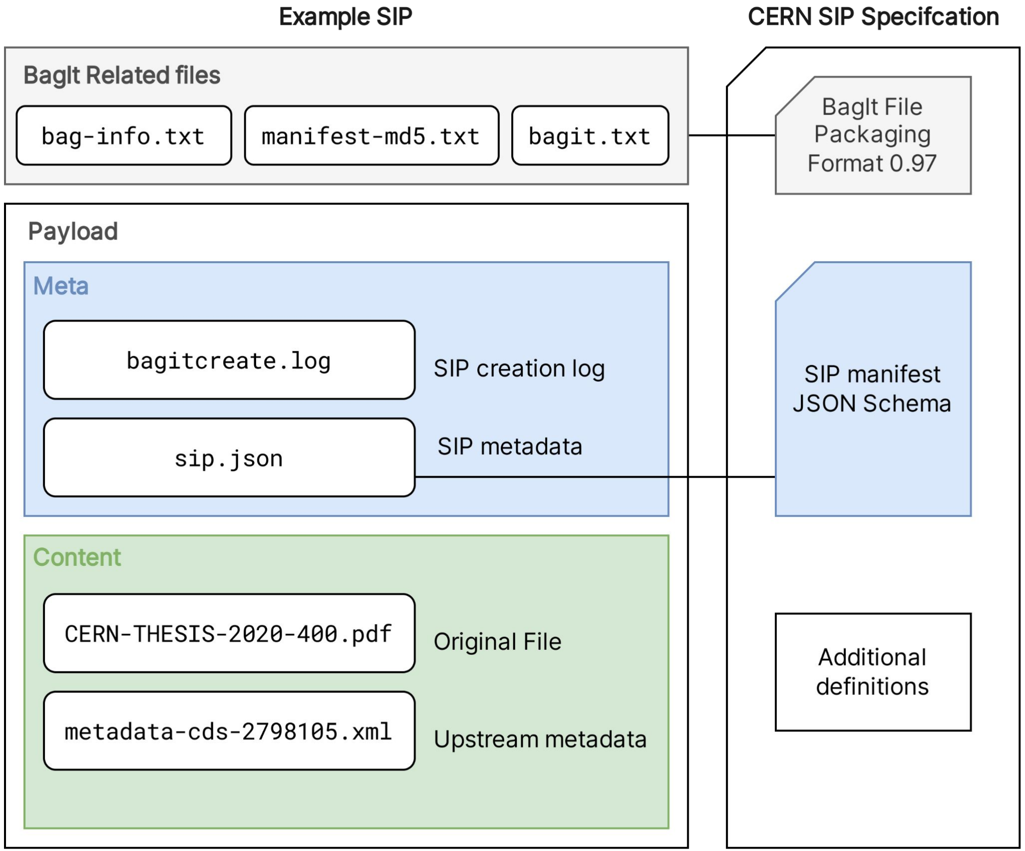
B. CERN users want to preserve their assets
   - released on digital repositories
   - local files

→ CERN Digital Preservation Strategy

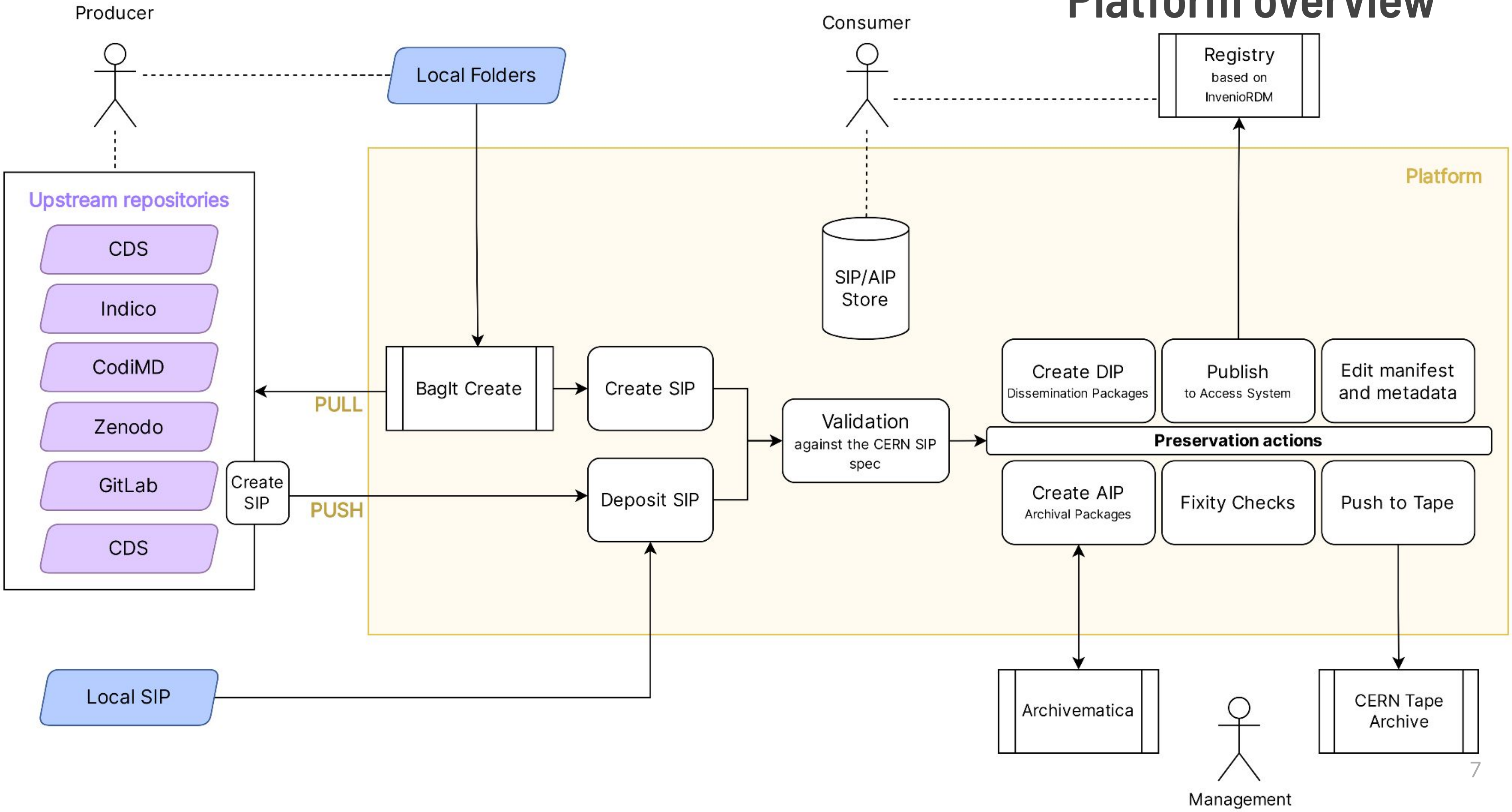
Creating SIPs

- **BagIt Create**: a tool to harvest data and export digital repository records in packages with a consistent format, according to a well defined specification
  - → CERN SIP Spec

- CLI or as a software package
  - `$ bic --source cds --recid 2748063`



Platform overview



Features

- SIP creation with BagIt Create
- AIP creation with Archivematica
- Push to Tape and Retrieve from Tape (CTA)
- (Optional) additional curation for local resources
- Fixity checks
- Dissemination and access to the archives

archivematica®

CERN Tape Archive

DIGITAL MEMORY REGISTRY

Technology

- Dev (and Git) Ops oriented approach to deployments
- Everything modular and OSS, with detailed documentation for usage and development
- CERN specifics documented and easily un-pluggable
- Platform: a Python Django restful web application
    - OpenAPI specs
    - Frontend in React
- Registry powered by InvenioRDM

Further improvements

- Moving SIP creation to the repositories
- Appraisal and content selection
- Archivematica and the support for Office documents
- Access to the Registry

References & Links

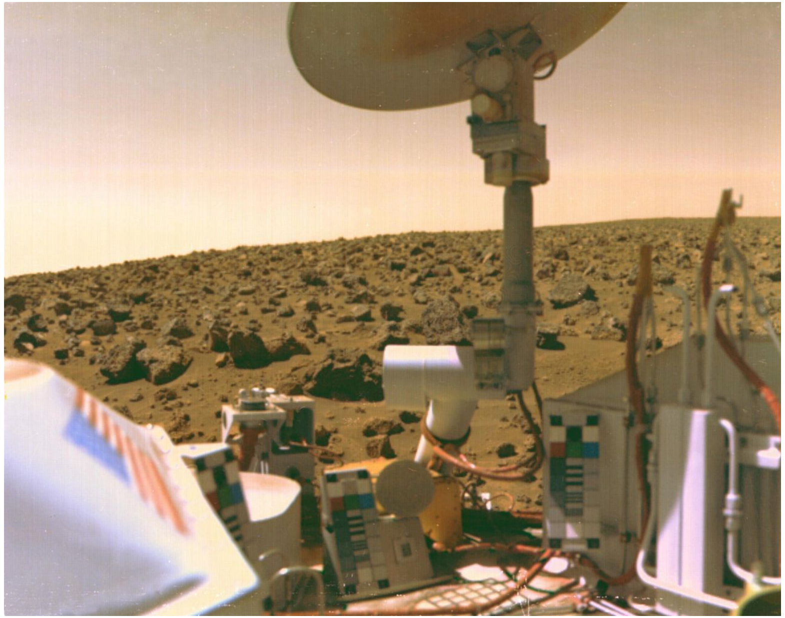
1. http://cds.cern.ch/record/2856775
2. https://gitlab.cern.ch/digitalmemory/bagit-create
3. https://gitlab.cern.ch/digitalmemory/oais-platform
4. https://gitlab.cern.ch/digitalmemory/sip-spec
5. https://wiki.archivematica.org/Format_policies
6. Paper: https://github.com/avivace/pv2023/releases/download/pv2023/PV2023-3.pdf

Backup

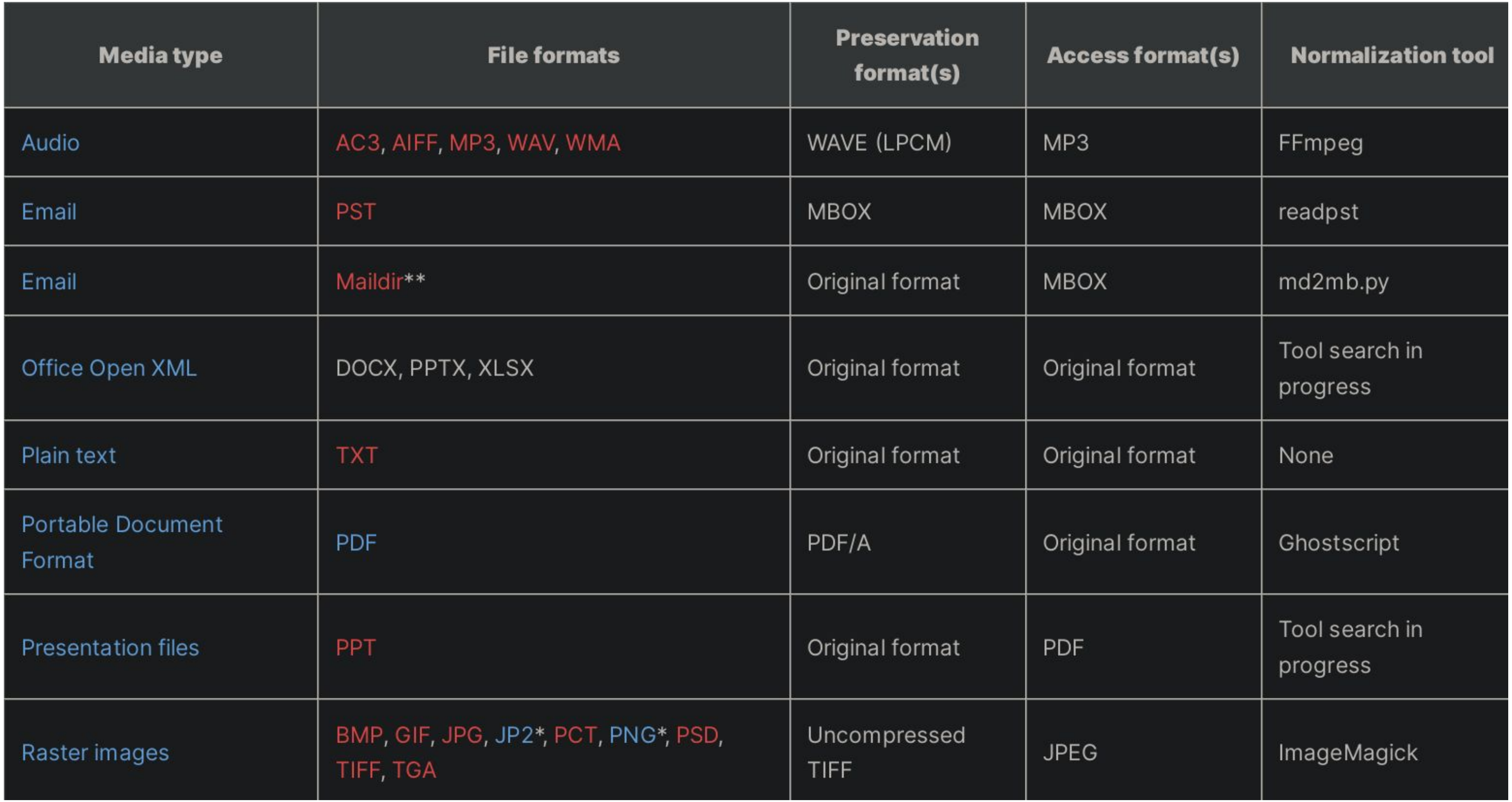
Risks & Challenges

1. Media which cannot be read
2. Information trapped in legacy systems
3. Incomplete metadata and uncomplete context
4. Unclear ownership & provenance
5. Corrupted or deleted files
6. Expired software licenses
7. Expired vendor supports
8. Lossy conversions or migrations

→ Digital Dark Age



Archivematica default format policies



Media type	File formats	Preservation format(s)	Access format(s)	Normalization tool
Audio	AC3, AIFF, MP3, WAV, WMA	WAVE (LPCM)	MP3	FFmpeg
Email	PST	MBOX	MBOX	readpst
Email	Maildir**	Original format	MBOX	md2mb.py
Office Open XML	DOCX, PPTX, XLSX	Original format	Original format	Tool search in progress
Plain text	TXT	Original format	Original format	None
Portable Document Format	PDF	PDF/A	Original format	Ghostscript
Presentation files	PPT	Original format	PDF	Tool search in progress
Raster images	BMP, GIF, JPG, JP2*, PCT, PNG*, PSD, TIFF, TGA	Uncompressed TIFF	JPEG	ImageMagick