

The challenge of Digital Preservation at CERN

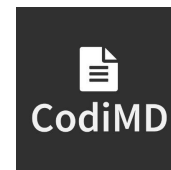


Contents

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



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

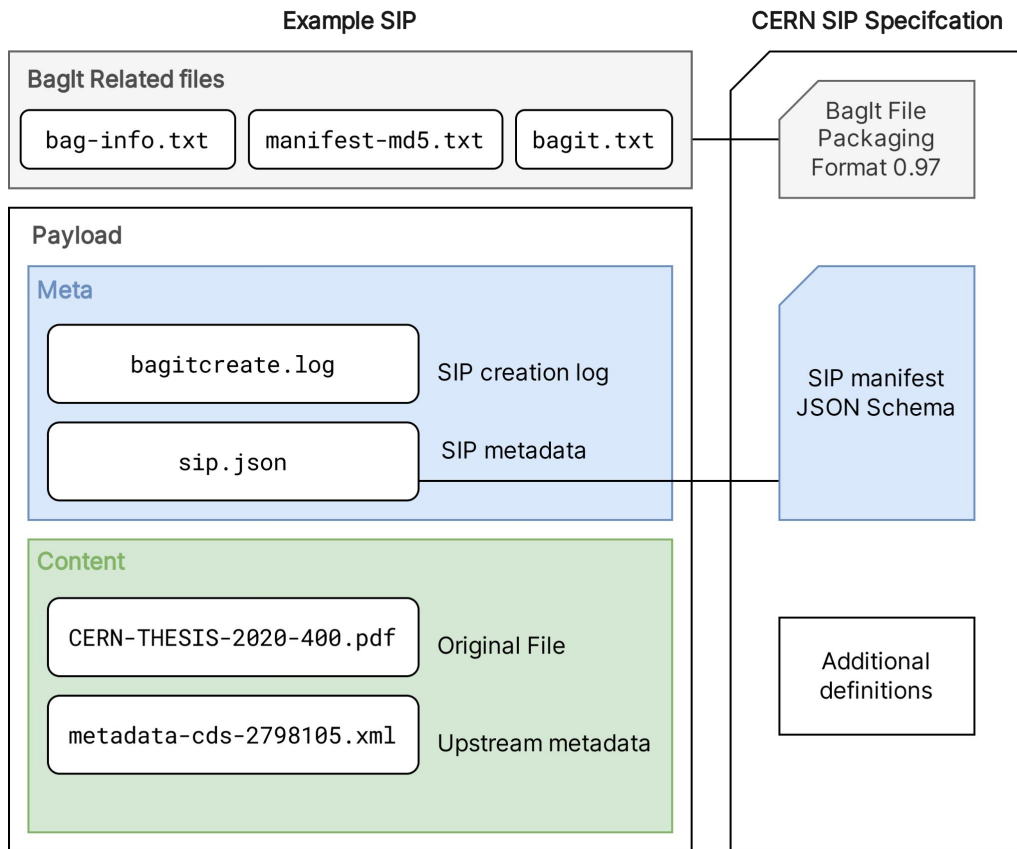
Scenarios

- 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

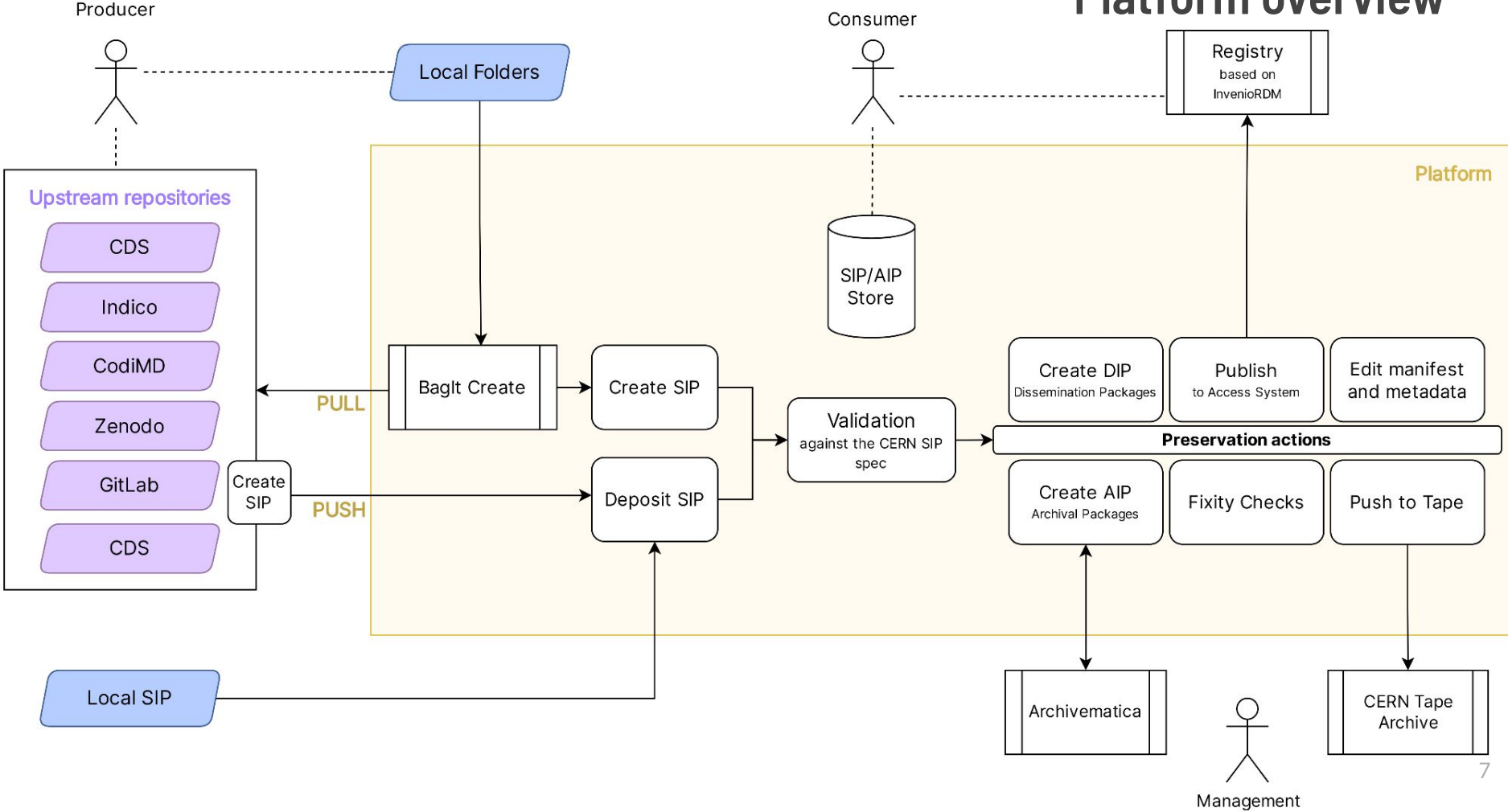
→ [CERN Digital Preservation Strategy](#)

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

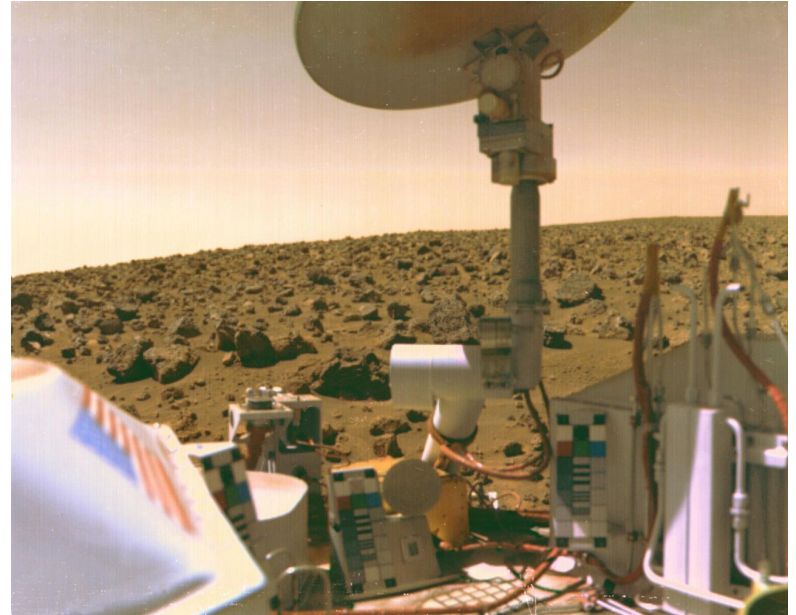
1. CERN Digital Preservation Strategy: proposal
<http://cds.cern.ch/record/2856775>
2. BagIt Create
<https://gitlab.cern.ch/digitalmemory/bagit-create>
3. OAIS Platform
<https://gitlab.cern.ch/digitalmemory/oais-platform>
4. CERN SIP Specification
<https://gitlab.cern.ch/digitalmemory/sip-spec>
5. Format Policies
https://wiki.archivematica.org/Format_policies
6. The Challenge of Digital Preservation at CERN
<https://cds.cern.ch/record/2857550>

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

Harvest UI

Harvest

Create SIP packages from the supported digital repositories (uses Bagit Create tool)

Query Search Record by ID **Source**

« < 1 > » Results per page:

Title	Record ID	Actions
Modernising the CERN CMS Trigger Rates Monitoring software	2798105	<input type="button" value="🌐"/> <input type="button" value="🗑"/>
CERN Summer Student Report: Digital Memory at CERN - Archiving and preserving existing knowledge	2779856	<input type="button" value="🌐"/> <input type="button" value="🗑"/>

Pipeline overview

General Archive Information

Record 37

Source: cds
ID: 2813809
Link: <https://cds.cern.ch/record/2813809>

Tags:

[EDIT MANIFEST](#)

Pipeline

The pipeline consists of five steps, each represented by a circle with a progress indicator:



- Harvest**: Completed
- Validate**: Completed
- Checksum**: Completed
- Invenio RDM**: Completed
- Upload to AM**: In progress

Steps



2

▼ Step Details



Registry results

Search uploads...  



42 result(s) found Sort by Recently updated ▾

 Draft June 1, 2020 (v1) Photo Metadata-only 



Preparing for HL-LHC: Increasing the LHCb software publication rate to CVMFS by an order of magnitude
No name
No description
Uploaded on June 29, 2022

 Draft June 1, 2020 (v1) Photo Metadata-only 

Development of a single-photon imaging detector with pixelated anode and integrated digital read-out
No name
No description
Uploaded on June 29, 2022

 Draft June 1, 2020 (v1) Photo Metadata-only 

AMS Highlights
No name
No description
Uploaded on June 29, 2022

 Draft June 1, 2020 (v1) Photo Metadata-only 

Registry example

CERN DIGITAL MEMORY Search records... Communities My dashboard + oais2@ce...

Published December 13, 2022 | Version 2, Archive 18

Publication Metadata-only

Digital Memory platform demo records

OAIS, Platform

Citation

Style APA

OAIS, P. (2022). Digital Memory platform demo records (2, Archive 18).

Description

Source: codimd
Link: <https://codimd.web.cern.ch/QzOjquDyTgudDIL5mw5aqQ#>

Additional details

OAIS Artifacts

SIP: [Download](#) (timestamp: 12/13/2022, 16:17:46, description: Submission Information Package as harvested by the platform from the upstream digital repository., path: /eos/user/o/oais/platform-storage/preserve/sips/sip::codimd::QzOjquDyTgudDIL5mw5aqQ::1670948266) , **AIP:** [Download](#) (timestamp: 12/13/2022, 16:19:51, description: Archival Information Package, as processed by Archivematica., path: /eos/user/o/oais/platform-storage/preserve/aips/ad98/8c79/f147/4e7a/a08c/41b0/49ea/9614/sip__codimd__QzOjquDyTgudDIL5mw5aqQ__1670948266__Archive_18-ad988c79-f147-4e7a-a08c-41b049ea9614.7z)

Created: December 14, 2022 Modified: December 14, 2022

[Edit](#)

[New version](#)

[Share](#)

Versions

Version 2, Archive 18	Dec 13, 2022
Version 1, Archive 18	Dec 13, 2022

[View all 2 versions](#)

Details

Resource type
Publication

Export

JSON [Export](#)

About InvenioRDM
Product page

Get involved
GitHub

Community
Chatroom