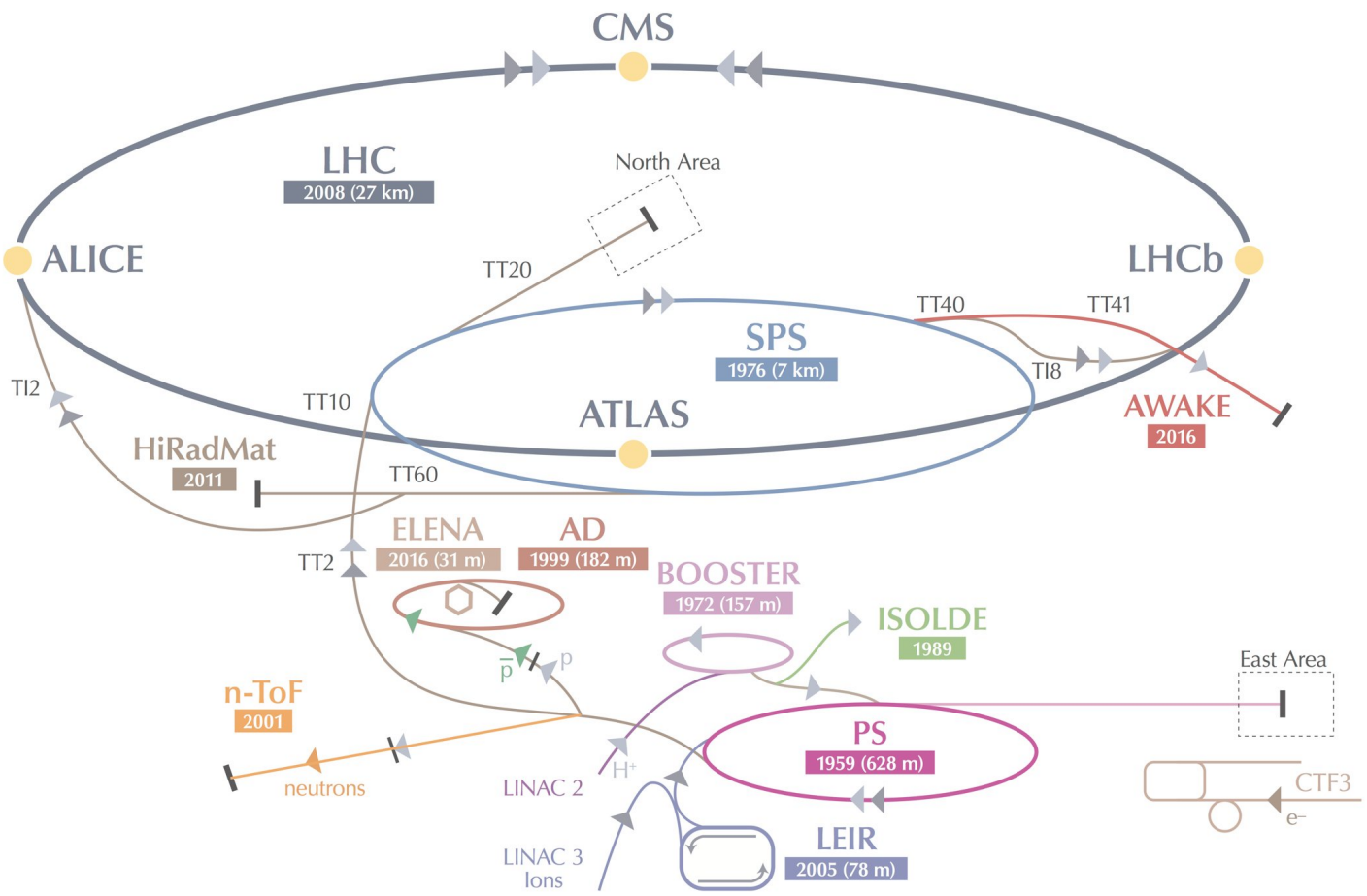
CERN’s Digital Memory

EIROforum

Knowledge Management Workshop 2024

JY Le Meur

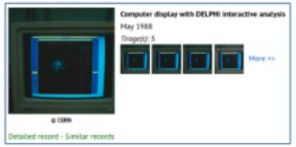

CMS
LHC
2008 (27 km)
North Area
ALICE
LHCb
TT20
TT40
TT41
SPS
1976 (7 km)
TI8
TI2
TT10
ATLAS
AWAKE
2016
HiRadMat
2011
TT60
ELENA
2016 (31 m)
AD
1999 (182 m)
TT2
BOOSTER
1972 (157 m)
ISOLDE
1989
East Area
n-ToF
2001
PS
1959 (628 m)
neutrons
LINAC 2
CTF3
LINAC 3
Ions
LEIR
2005 (78 m)

CERN Digital Memory

From preservation by chance to preservation by mission

Rescue operations

- 20th century **analog** multimedia was entirely digitized into preservation formats



- **Digitally-born** data hosted in phased out systems was converted into **A**rchival **I**nformation **P**ackages stored on CERN Cloud (e.g. the ILC document server)

Prevention actions

WORK IN PROGRESS

Adhering to the Open Archival Information System (**OAIS**) model

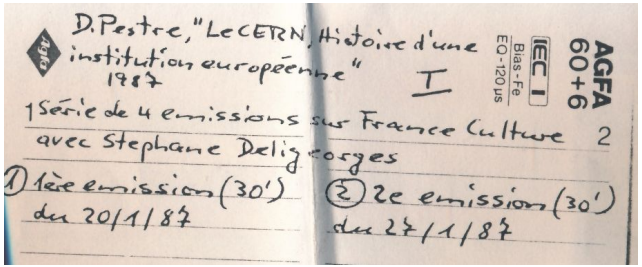
- Providing Information Systems with solutions to create preserved bags (AIP)
- Providing Users with interfaces to select data to be digitally preserved

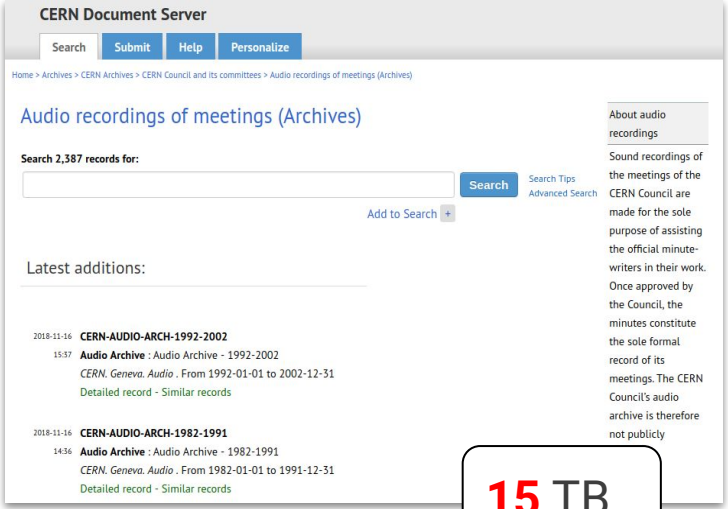
AUDIO

Treated: 5112 cassettes, 3289 tapes

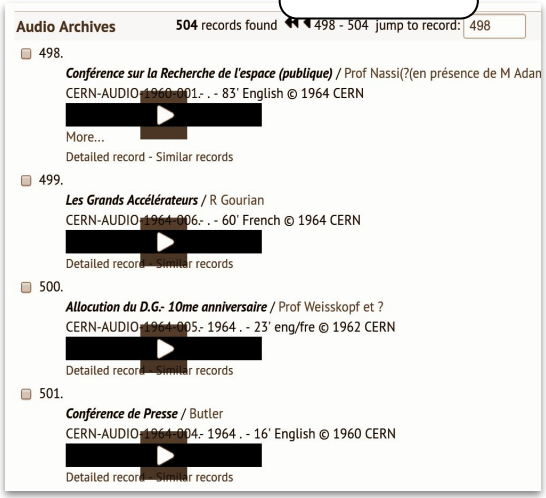
1’740 Kg

~30’500 audio files loaded into **3’000 records** - & **3’300 scanned timelines**





15 TB



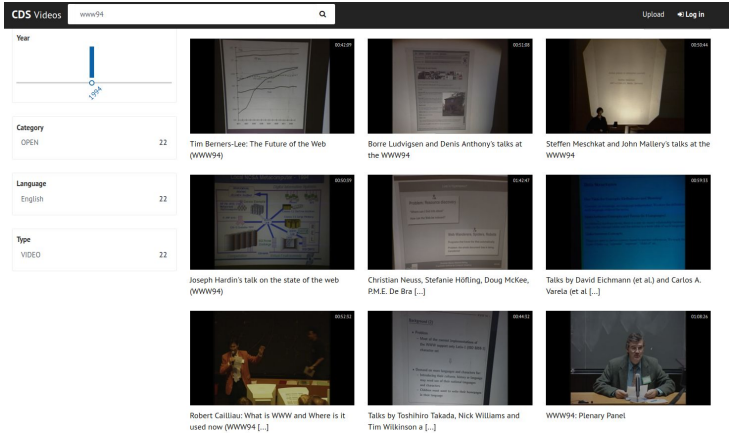
From Data to knowledge ?

- Get indexed text with speech to text / OCR
- Merge with digitally-born audios
- Open access → many decades embargo

VIDEO

	Total
35 mm	4
16 mm	112
8 mm	5
D3	62
Digital Betacar	5
DVCAM	446
DV/miniDV	151
DVD	676
Betacam SP	1368
Betacam	33
1 inch C	21
U-matic S/SP	1066
U-matic	199
VHS	911
Others	44
Total	5109

Documentaries, movies, conferences, clips, lectures, footages…



https://videos.cern.ch/search?q=www94

Originals		Preservation Master	Access Master	Access Copy
35mm films	Wrapper:	.mkv	.mov	.mp4
	Video codec:	FFV1 – 10 bits RGB	Apple ProRes 422 LT	H.264 @ 5Mbps
	Audio codec:	24 bits PCM, 48kHz	24 bits PCM, 48kHz	16 bits AAC, 44.1kHz, 256kb
	Definition / Aspect ratio:	4096x? / Original	1920x1080 / Pillar-letterbox	1920x1080 / Pillar-letterbox
16mm films	Wrapper:	.mkv	.mov	.mp4
	Video codec:	FFV1 – 10 bits RGB	Apple ProRes 422 LT	H.264 @ 5Mbps
	Audio codec:	24 bits PCM, 48kHz	24 bits PCM, 48kHz	16 bits AAC, 44.1kHz, 256kb
	Definition / Aspect ratio:	2048x? / Original	1920x1080 / Pillar-letterbox	1920x1080 / Pillar-letterbox
Analogue and digital SD video	Wrapper:	.mkv	.mov	.mp4
	Video codec:	FFV1 – 10 bits YCbCr	Apple ProRes 422 LT - SD profile	H.264 @ 1Mbps
	Audio codec:	24 bits PCM, 48kHz	24 bits PCM, 48kHz	16 bits AAC, 44.1kHz, 256kbps
	Definition / Aspect ratio:	?x576 / Original	?x576 / Original	640x360 / Pillar-letterbox

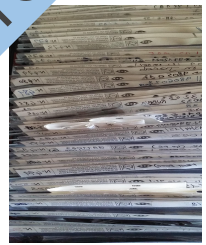
15K hours

~100 TB

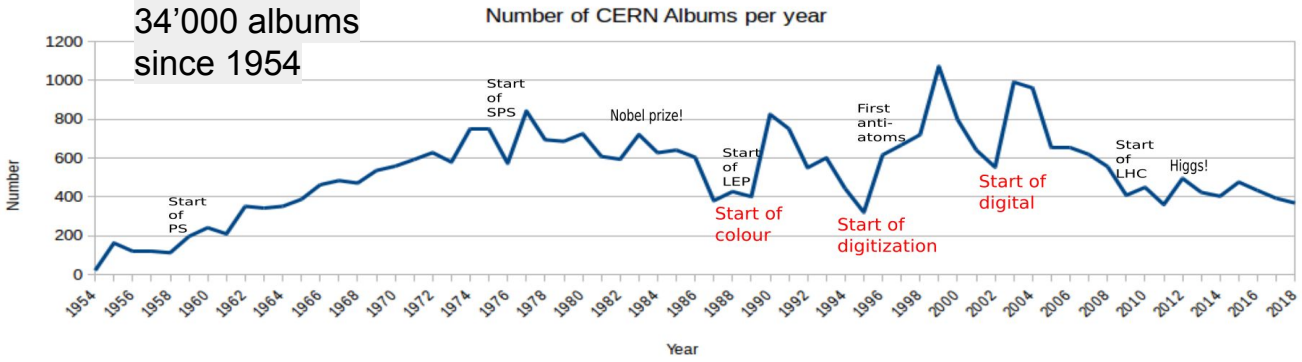
Challenges

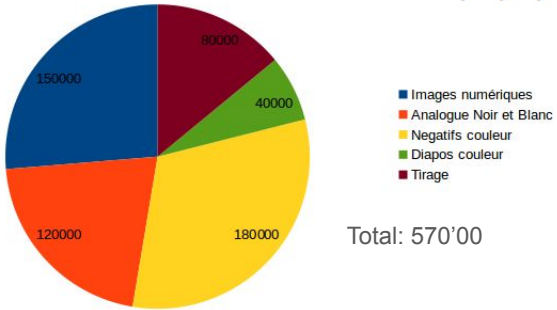
- Sort out: classify, deduplicate, privacy issues, access control
- Tricky automated transcription
- Complex curation work
- Very high value
- Unknown future (formats)

PHOTO

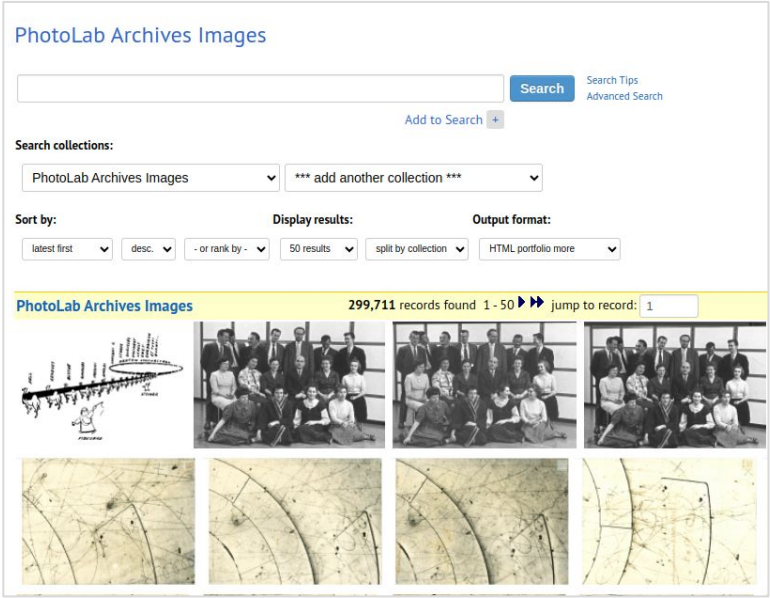


34’000 albums since 1954





Total: 570’00



Next steps

- The captioning challenge
- The retiree network
- The Bubble chambers images

Digitizing is only the first step to expand the lifetime of obsolete data.

Preserving with FAIRness is the next challenge.

Targeted information systems

Main Live digital repos used at CERN

- OPEN DATA, scientific datasets
- Multimedia repo
- CDS Institutional repo
- INSPIRE, disciplinary repo
- EDMS/PLM, engineering repo
- INDICO, event management
- GITLAB, code repo
- ZENODO, world-wide repo
- Other info systems : Admin (EDH), Drafts, Wikis, Social Media, Emails, Web sites, etc.

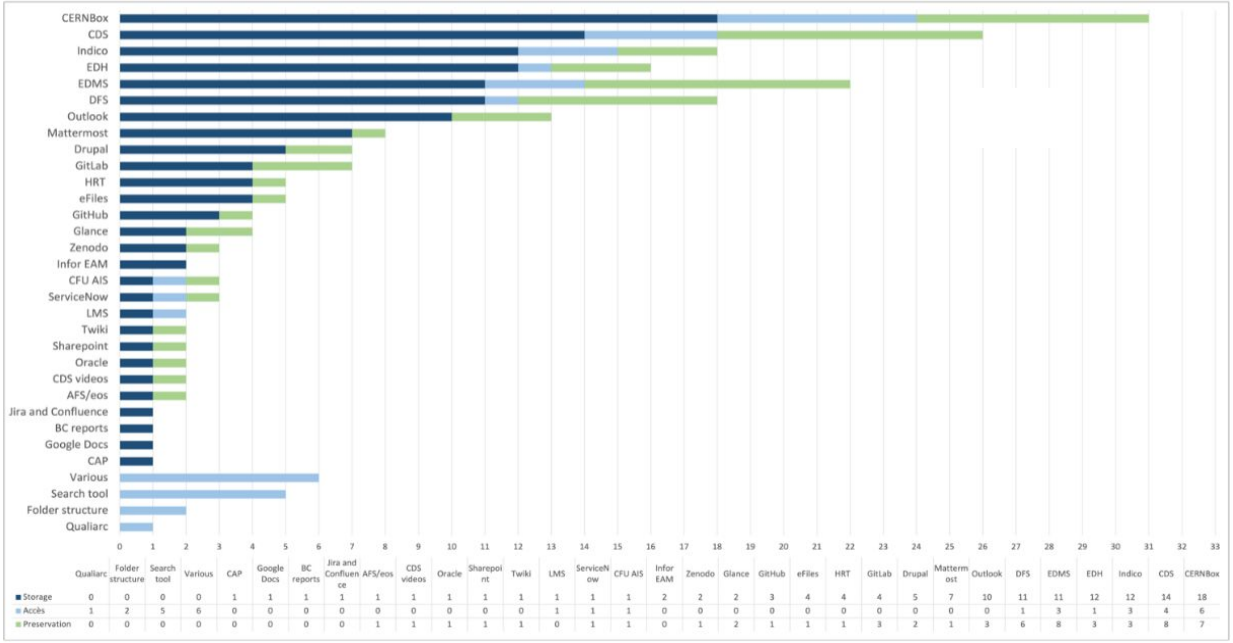
…

Local folders (user-stored content)

- E.g. Slides submitted to external conferences, notes, drafts, etc

Record Management

Tools used for accessing, searching, storing, preserving and managing retention periods of documents



	Qualiarc	Folder structure	Search tool	Various	CAP	Google Docs	BC reports	Jira and Confluence	AFS/eos	CDS videos	Oracle	Sharepoint	Twiki	LMS	ServiceNow	CFU AIS	Infor EAM	Zenodo	Glance	GitHub	eFiles	HRT	GitLab	Drupal	Mattermost	Outlook	DFS	EDMS	EDH	Indico	CDS	CERNBox
Storage	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	3	4	4	4	5	7	10	11	11	12	12	14	18
Accès	1	2	5	6	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	3	2	1	3	6	8	3	3	8	7
Preservation	0	0	0	0	0	0	0	0	1	1	1	1	1	0	1	1	0	1	2	1	1	1	3	2	1	3	6	8	3	3	8	7

Crédit: Salomé Rohr (SIS Survey May 2024)

Department Records Officers (DRO) network

- 1 per department and per experiment
- ~20 people mandated to support record management and archiving plan

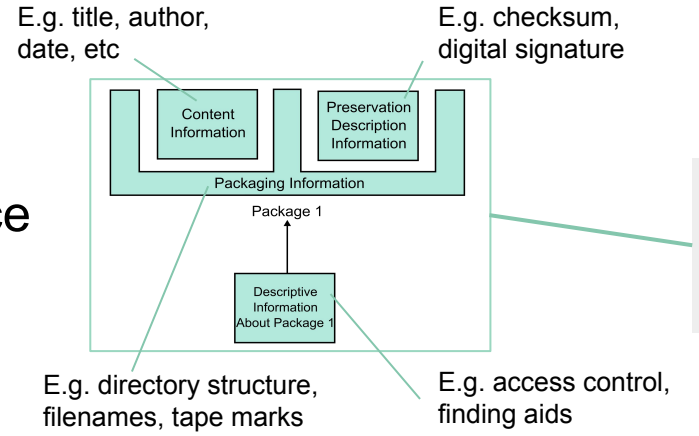
Document types	Number of mentions/22
Reports	17
Presentations/slides	17
Technical documentation	16
Memoranda	15
E-mails	14
Meeting minutes	14
Financial files	13
Photographs	12
Scientific papers	11

Digital Memory: PReservation As A Service

- Establish a **Preservation Policy**
  - Collaboration with the central library
  - Scope, rules and responsibilities

The **OAIS** reference model rules how preservation should be applied - ISO/Seal to be a Trusted Digital Repo

- Empower CERN repos with **Preservation Features**
  - Simple workflows between repos and preservation service
  - Help getting certifications required by funding agencies



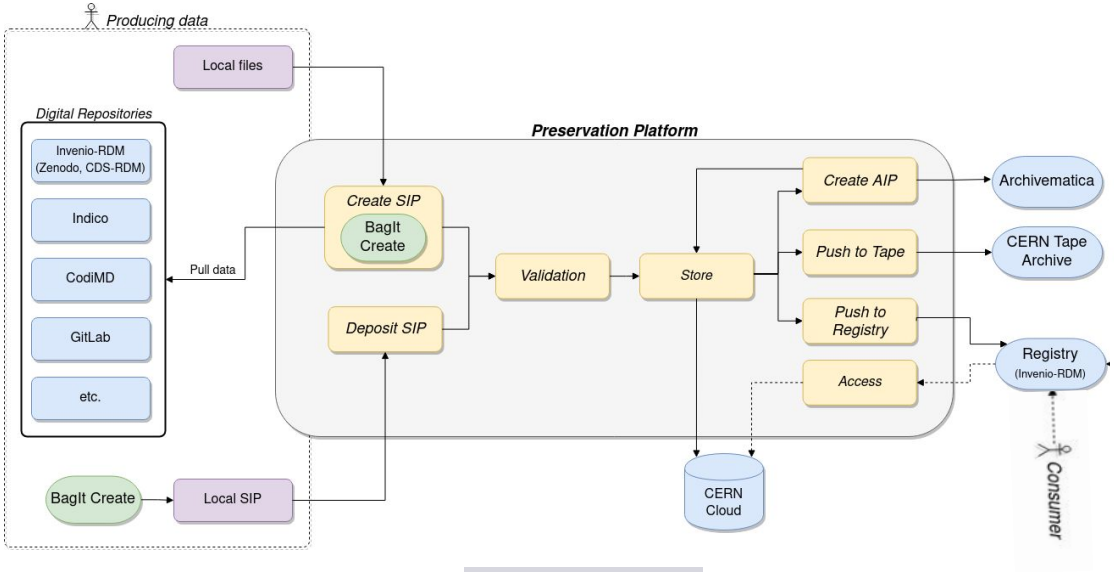
Preserved Bag aka AIP

Agreement between Digital Repos and Preservation Service

1. Repo ask Preserve platform to harvest their resources
2. Repo submits Packages (SIPs) to Preserve platform
3. Repo registers Preserved Bags (AIPs) to Preserve Registry

CERN users actions:

A. Release data on the Repos (mandatory)
B. Load local files to CERN **Preservation Platform** (via Web portal; optional)



scenarios 1 & 2

Welcome to Preserve, the CERN Digital Memory Platform

This service allows to retrieve your digital assets from CERN digital repositories, request their archival and prepare them for long term preservation, complying to the OAIS specifications.

Features overview

1 — Submit

Users can trigger harvesting of records from supported CERN digital repositories.

SUBMIT

2 — Preserve

The platform will prepare the data for long term preservation, creating Archival and Dissemination packages and saving them on the CERN Tape Archive.

PRESERVE

3 — Access

Processed data is finally recorded in our Preservation Registry, where it is available to the authorized people.

ACCESS

Harvest Data

Here's how to get started and make the platform retrieve records from CERN supported repositories for you.

1 — Configure

To be able to fetch your private records you'll need to set up some API tokens first.

CONFIGURE

2 — Search

You can search for any records on Indico, CodiMD, CDS and Zenodo from the Search page. The platform will fetch the data for you.

SEARCH

3 — Organize

Once you selected your records, they will be waiting in the 'Staging Area', where you can organize them with custom tags.

ORGANIZE

4 — Download

Once you confirmed your selection from the staging area, the preservation process will start. Check after a while to get download links.

DOWNLOAD

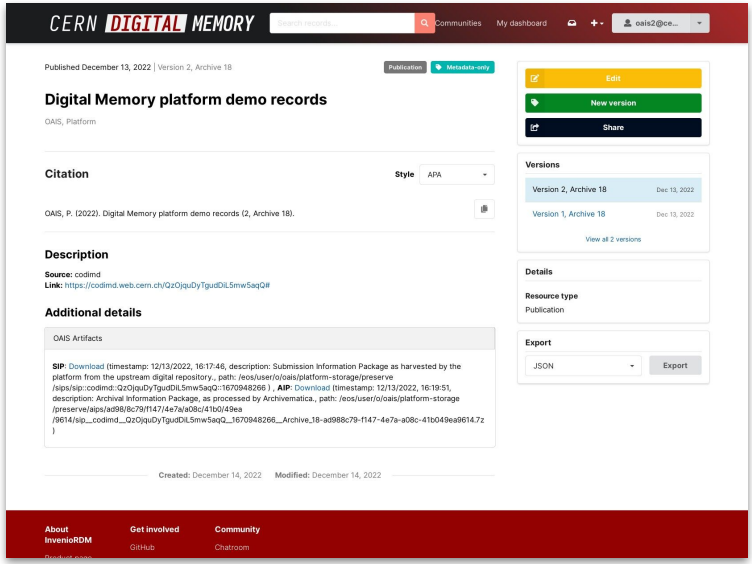
Sources availability

Here's an overview of the available repositories to harvest from. Some of them may need additional configuration.

Source	Configuration Status
CERN Document Server	Only public records will be available. Configuration is needed for private records.
CodiMD	Source unavailable. Additional configuration is needed.
Indico	Source ready
Zenodo	Source ready

Support for additional repositories is work in progress.

http://preserve.web.cern.ch



Add resource

Here you can find different ways to import data into the platform to start its long term preservation process.

Search and harvest

Search for records and documents from various CERN digital repositories (e.g. CDS, Invenio, Indico) and let the platform harvest the record for you.

Query

E.g. "REANA data analyses"

Search Record by ID

Source

cds

- cds
- indico
- codimd
- zenodo

SEARCH

FIND ALL YOUR RECORDS ON CDS

FIND ALL YOUR EVENTS AND CONTRIBUTIONS ON INDICO

Harvest from URL

Enter a URL from the supported digital repositories (CDS, Invenio, CERN opendata) and we'll try to find the record for you.

URL

https://cds.cern.ch/record/2798105/

PARSE URL

Upload folder

Upload files and folders from your local machine.

Select folder: Browse... No directory selected.

UPLOAD

Advanced features

Here are some more advanced workflows to submit your data.

Upload Submission Bag

Upload a Submission Bag from your local machine (as a ZIP file). To create such bags you can use the BagIt Create tool or check the format specification.

Select compressed SIP: Browse... No file selected.

UPLOAD

Announce Submission Bag

If you already uploaded your SIP on EOS, you can add it to the platform by entering its absolute path here. Make sure you have granted the necessary permissions (give the "oais" user read access if the folder is private) and that the path directly points to the SIP folder (i.e. it contains `data/meta/sip.json`).

EOS Path *

/eos/home-u/user/sip_folder

ANNOUNCE

Batch Announce Submission Bag Folders

If you already uploaded your SIPs on EOS, you can add it to the platform by entering the parent folder's absolute path here. Make sure you have granted the necessary permissions (give the "oais" user read access if the folder is private) and that the path directly points to the folder containing the SIP folders (i.e. it contains SIP folders).

EOS Path *

/eos/home-u/user/sip_parent_folder

Tag *

ilcdoc_2023

ANNOUNCE

Archives

This page shows the list of your archives. You can browse through the created archives and get more details.

Filter

Package state: AIP

Source: cds
- cds
- indico
- codimd
- zenodo

Tag: ilcdoc200-300

FILTER

Last Step: Step

Filters used: ilcdoc200-300

SELECT ALL ON PAGE or SELECT ALL 100 RESULTS or DESELECT ALL

ACTION ON 0 SELECTED

Select	Last Update	Original Record	Title	State	Last Step	
☐	2024/06/04 11:06:51	16368 (ilcdoc)	ilcdoc - 16368	AIP ✓	Upload to AM ✓	DETAILS
☐	2024/06/04 11:06:51	16364 (ilcdoc)	ilcdoc - 16364	AIP ✓	Upload to AM ✓	DETAILS
☐	2024/06/04 11:05:51	16366 (ilcdoc)	ilcdoc - 16366	AIP ✓	Upload to AM ✓	DETAILS
☐	2024/06/04 11:05:51	16372 (ilcdoc)	ilcdoc - 16372	AIP ✓	Upload to AM ✓	DETAILS

General information

ilcdoc - 16368

EDIT MANIFEST

Record ID: 940
Source: ilcdoc
ID: 16368
Link: N/A
Tags: ilcdoc200-300

Choose Next Step
- UPLOAD TO AM
- PUSH TO REGISTRY
- PUSH TO TAPE

Pipeline overview

Announce Completed → Validate Completed → Checksum Completed → Upload to AM Completed → +

Steps

Announce

Artifacts: SIP

Step Details

ID: 2886 **Start Date:** 2024/05/27 16:02:13 **End Date:** 2024/05/27 16:02:14 **Status:** Completed

Celery Task Details

Output data

Validate

Step Details

ID: 2891 **Start Date:** 2024/05/27 16:02:14 **End Date:** 2024/05/27 16:02:15 **Status:** Completed

Celery Task Details

Output data

Checksum

Step Details

ID: 2894 **Start Date:** 2024/05/27 16:02:15 **End Date:** 2024/05/27 16:02:15 **Status:** Completed

Conclusion

- Digital Memory project is aiming at preserving knowledge across generations
- First action was to prevent the decay of analog multimedia material
  - Large-scale digitization of past audio and images
- Focus on the best world-wide practices in digital preservation
  - Aligning with the OAIS reference model: policies and technology
- New layer on top of existing Information repositories
- Set up a Preservation platform
  - Central hub supporting repositories and users

“CERN is not just another laboratory. It is an institution that has been entrusted with a noble mission which it must fulfil not just for tomorrow but for the eternal history of modern thought.”

Albert Picot, 3rd session of CERN Council, Geneva, 10 June 1955

Thanks for your attention !