

### **ARCHIVER** Archiving and Preservation for Research Environments

**EOSC Symposium - ARCHIVER Panel** 

Speaker: João Fernandes (CERN)

Panelists: Hervé l'Hours (LTDP TF), Matthew Addis (Arkivum), Teo Redondo (Libnova)

14th November 2022



ARCHIVER - Archiving and Preservation for Research Environments project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824516.



### **ARCHIVER Project**

Focus: Archiving and Data Preservation Services using cloud services available via the European Open Science Cloud (EOSC) Procurement R&D budget: 3.4M euro; Total Budget: 4.8M meosc Starting Date: 1<sup>st</sup> of January 2019 **Duration: 42 Months <u>Coordinator</u>: CERN (Lead Procurer) European Commission** EMBL-EBI PIC port d'informació científica DESY. ) Buyers Group (BG) - Public organisations committing funds to contribute to a joint-R&D-procurement, research data use cases and destine **Buyers** R&D testing effort. Consortium HIGH ENERG PHOTON-NEUTRON ASTROPHYSICS | LIFE LONG TAIL C SCIENCES COSMOLOGY SCIENCES PHYSICS SCIENCE **Experts** addestine **Experts** - Partner organisations bringing coper tise in requirement assessment and promotion activities



### **Problem and Solution**



k ARCHIVER "Current state of the art" Report

	EMBL 2 – Clo	PIC 1 – Large	PIC 2 – Mix F	PIC 3 – Data	CERN 2 – CE	CERN 3 – CE	
pr	oject.e	eu/deplo	oyment-	-scenar	ios		
<u>)</u> S:	://doi.o	org/10.52	281/zer	nodo.36	<u>618215</u>		

**Distribution** 

PIC port d'informació científica

File Storage

Storage

Caching

р

FIRE

<u>\_</u>

EMBL



ER

Open Data

Z

High level services: visual representation of data (domain specific), reproducibility of scientific analyses, etc.

User services: search, discover, share, indexing, data removal, etc. Access under Federated IAM

OAIS conformant services: data readability formats, normalization, obsolesce monitoring, files fixity, authenticity checks, etc. ISO 14721/16363, 26324 and related standards

Layer 1 Storage/Basic Archiving/Secure backup

**R&D** - Scope

Layer 4

Advanced

services

Layer 3

**Baseline user services** 

Layer 2

Preservation

Data integrity/security; cloud/hybrid deployment Data volume in the PB range; high, sustained ingest data rates in Gb/s. ISO certification: 27000, 27040, 19086 and related standards. Archives connected to the GEANT network

# **Digital Memory** Individual Scientist RN

DES

Scientific use cases deployments documented at: https://www.archiver-

ARCHIVER "current state of the art" report in the context of the EOSC: http

EMBL

2

DESY

Experiment

EUXFEL

3

DESY



### **Project Timeline**





### **ARCHIVER Resulting Services**



https://archiver-project.eu/archiver-long-term-data-preservation-solutions



### **Resulting Services Sustainability**

 Total Cost of Service (TCO) modelling calculators: ability to optimise cost considering volumes, access frequencies, safety, processing and retention periods to achieve economic sustainability





 Comprehensive mapping against CoreTrustSeal, DPC RAM, etc. as service providers, supporting the implementation of good practice for LTDP of Data Stewards



### **ARCHIVER Award Winning Project**



The International Council of Archives Award for Collaboration and Cooperation





### **ARCHIVER White Paper**

ARCHIVER Consort	ium (CERN, EMBL-EBI, DES)	(PIC, Addestino & TRUST-IT)	
CERNY	EMBL-EBI	DESY	
No provent	addestine	Trust IT Services	
& ARCHIVER Pilot P	hase Lead Contractors (A	rkivum, Libnova)	
	libn	ova	
11			
			•
			•
	•		

#### Disclaimer

ARCHIVER project with Grant Agreement number 824516 is a Pre-Commercial Procurement Action funded by the EU Framework Programme for Research and Innovation Horizon 2020.

This document contains information on the ARCHIVER core activities, findings, and outcomes, and it may also contain contributions from distinguished experts who contributed to ARCHIVER. Any reference to content in this document should clearly indicate the authors, source, organisation, and publication date. This document has been produced as a result of a project co-funded from the European Commission. The content of this publication is the sole responsibility of the ARCHIVER consortium and selected Contractors of the Pilot Phase and cannot be considered to reflect the views of the European Commission.

- ARCHIVER services Technical Deep Dive
- Financial and Environmental Sustainability
- ARCHIVER R&D Methodology
- Future R&D opportunities

#### **Publication in Zenodo mid-December**



ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

# **Thank You!**



ARCHIVER - Archiving and Preservation for Research Environments project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824516.



### **Panel Questions**

What were the main challenges ARCHIVER wanted to address and how were they overcome? Were they organisational, technological, economic or a mix of these?

What are the three most important features of each of the ARCHIVER resulting solutions?

Research data lives for longer than any vendor, system or technology. How do the ARCHIVER resulting LTDP services prevent vendor lock-in and encourage portability and interoperability, yet at the same time make it attractive for commercial services to participate? Are these in conflict? What happens when contracts end?

What are the TCO calculators implications with CoreTrustSeal: how do the capabilities required by CTS requirements influence the costing factors of the services proposed by the ARCHIVER resulting services? Can the TCO calculators be distributed in the EOSC context?

How can LTDP services support research data producers in aspects such as to preserve and re-run complete software environments, beyond the original software applications, ranging from the large scale through to the long tail, in the context of the EOSC?



### **Additional Slides**



ARCHIVER - Archiving and Preservation for Research Environments project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824516.



### Selected Consortia: Arkivum

- SaaS solution in the cloud or on-premise, provided as a managed service
- Autoscaling based on Kubernetes, supporting multi-petabyte volumes of billions of objects
- Certified ISO 9001 / ISO 27001
- Solution supports NDSA Levels of Preservation, DPC RAM, CoreTrustSeal and FAIR principles, with mappings available.
- Integrations with RDM applications and FAIR environments



Prototype architecture of the Arkivum consortium (image courtesy of the Arkivum consortium)











XRootD 🛠eduGAIN





### Selected Consortia: Arkivum



### INVENIORDM .....

My dashboard

+-

#### Preview

You are previewing a new record that has not yet been published.

on	Home > C_test_data/CERN/HiggsToBBNtupieProducerTool/ > O_test_data/CERN/H				es	
ashboard + dmin +	100 etc	···· + Q search	Published 2019   Version v1			
es –	Content Type	а	for Hbb tagging ML studies	,	1000	
olorer		ntuple_merged_0.h5 /arkivum/higge1/HiggsToBBNtupieProducerTool/Test/Intuple_merged	HiggsToBBNTuple_HiggsToB	Version v1	2019	
xorts +	-	ntuple_merged_1\h5. /arkivum/higgsi\HiggsToBBKtupleProducerToot/Test/ntuple_merged	Duarte, Javier	Details		
Rules +	-	ntuple_merged_2.h5 /orkivum/higgs1/HiggsToBBNtupleProducerTool/Test/ntuple_merged,	Citation	Resource type Dataset Publisher CERN Open Data Portal		
hual ingest tifications	m	ntuple_merged_3.h5 /arkivum/higgs1/HiggsToBBNtupleProducerTool/Test/ntuple_merged,	Duarte, J. (2019). Sample with jet, track and secondary vertex			
	m	ntuple_merged_4.h5 /arkivum/higgsi/HiggsToBBNtupleProducerTool/Test/ntuple_merged,	ML studies HiggsToBBNTuple_HiggsToBB_QCD_RunII_13Te Data Portal.			
	-	ntuple_merged_5:h5 /ark/vum/higgs1/HiggsTaBBNtupleProducerTapl/Test/htuple_merged,			Rights	
		ntuple_merged_6.h5 /arkivum/higgs1/HiggsToBBNtupleProducerTool/Test/ntuple_merged.	Description The dataset consists of particle jets extracted from simulated p	Creative Commons Zero v1.0 Universal		
		ntuple_merged_7.h5 /orkivum/higgsi/HiggsToBBNtupieProducerTool/Test/ntuple_merged.	13 TeV generated with Pythia 8. It has been produced for deve originating from a Higgs boson decaying to a bottom quark-an guarker absorbed wards (2020) multiple and values.			
		ntuple_merged_8.h5 /orkivum/higgel/HiggsToBBNtupleProducerTool/Test/htuple_merged	The reconstructed jets are clustered using the anti-KT algorithm with R=0.8 from particle flow (PF) candidates (AK8 jets). The standard L1+L2+L3+residual jet energy corrections are applied to the jets and pileup contamination is mitigated using the charged hadron subtraction (CHS) algorithm. Features of the AK8 jets with transverse momentum pT > 200 GeV and pseudorapidity InI < 2.4 are provided. Selected features of inclusive (both charged and neutral) PF candidates with pT >		Export	
		files.txt /orkivum/higgs1/HiggsToBBNtupieProducerTool/Train/files.txt			JSON +	Export
	<u></u>	ntuple_merged_10.h5 /orkivum/higgs1/HiggsToBBNtupieProducerTool/Train/ntuple_mergec	0.95 GeV associated to the AK8 jet are provided. Additional fe charged particle track) with pT > 0.95 GeV associated to the A			
	<u></u>	ntuple_merged_11.h5 /ork/vum/higgs1/HiggsToBBNtupieProducerTool/Train/ntuple_merged_11.h	s /arkivum/higgsl/Hi +++			
	<u> </u>	ntuple_merged_12.h5 /arkivum/higgsi/HiggsTaB8NtupieProducerTool/Train/ntuple_merged_12/	15 /arkivum/higgst/Hi	Example of dataset landin	na pages created in Ir	venio
		ntuple_merged_13.h5 /arkivum/higast/higast/a88NtupleProducerTool/Train/ntuple_merged_13.	ns /arkivum/higgst/Hi ····	for research datasets i	n the Arkivum LTDP	solutic

## Selected Consortia: LIBNOVA

- Complete re-engineering of the service during ARCHIVER resulting in a new product line: LIBNOVA LABDRIVE
- Used infrastructure provided by AWS to validate unmet scalability; can be deployed on-premises;
- Running on Kubernetes with an adjustable number of components based on service demand which translates to cost and environmental effectiveness.
- Developed a generic reproducibility engine, now integrated in the new line of products, supporting a variety of applications (scientific workflows to obsolescent software applications)
- Certified ISO 27001/27018/27017
- Mappings against CoreTrustSeal; Audits against ISO16363 with processes and procedures based on the new resulting LABDRIVE product





## C Libnova



The platform has been made capable of **ingesting 15 Petabytes of content in 30 days**, at the remarkable performance of **500TB/day**, with low overall costs and environmental impact.

This includes fixity generation, characterization, validation, AV scan, technical metadata extraction, full text indexing, multiple copies generation and integrity verification.

líЬ



### Selected Consortia: LIBNOVA

- New generic reproducibility engine developed from LIBNOVA
  - R&D during ARCHIVER using REANA as an example
  - Engine generalised to support radically different needs: Jupyter notebooks, obsolete software (Lotus123, old games, etc.)

