



ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

# PILOT PHASE KICK-OFF EVENT AND AWARD CEREMONY

29 November 2021

Contact: [info@archiver-project.eu](mailto:info@archiver-project.eu)

Project website: [www.archiver-project.eu](http://www.archiver-project.eu)



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.

# Welcome to the Pilot phase!

Started in January 2019, ARCHIVER is a unique initiative currently running in the EOSC framework that is competitively procuring R&D services for archiving and digital preservation for data generated in the petabyte range.

Between December 2020 and August 2021 three consortia worked on innovative, prototype solutions for Long-term data preservation, in close collaboration with CERN, EMBL-EBI, DESY and PIC.

# Public Award Ceremony

The selection process for proceeding to the next phase is now over and the consortia selected to continue with the pilot phase will be officially announced on the public ceremony today!



## Event Outline

14:30 - 14:40: Welcome from Sergey Yakubov (DESY)

14:40 - 15:00: Project overview / update - João Fernandes (CERN)

15:00 - 15:30: Expected outcomes of the Pilot Phase - Buyers Group representatives (CERN, DESY, EMBL-EBI, PIC)

15:30 - 15:40: Interactive brainstorming

*15:40 - 15:50: Break*

## Award ceremony

15:50 - 16:20: Presentation from consortium 1

16:20 - 16:50: Presentation from consortium 2

16:50 - 17:00: Closing remarks - Sergey Yakubov (DESY)





## HOUSE KEEPING

This event is being recorded in its entirety. A link to the full recordings will be shared with participants afterwards

Two Q&A sessions are foreseen before the break and towards the end: keep your questions and use the chat!

Please don't activate your microphone and videos unless the host gives you permission



ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS



Deutsches Elektronen-Synchrotron - DESY  
Ein Forschungszentrum der Helmholtz-Gemeinschaft

Welcome!

Sergey Yakubov – DESY



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.



# DESY

## *"German Electron Synchrotron"*

**Founded** in 1959 as Germany's national accelerator laboratory (appointment to U Hamburg of Prof. Jentschke)

**Location** Hamburg; since 1992 second site in Zeuthen (close to Berlin)

Currently ~2500 staff;  
~3000 visitors p.a.

4 research divisions:  
particle physics, astro-  
particle physics, photon  
science, accelerator physics



## Mission

We conduct top-level international research into the fundamental relationships of matter – its structure and function. We are creating the knowledge base that is needed in order to solve the huge and urgent challenges that are facing society, science and the economy. The research facilities we develop and operate for this purpose are open to scientists from all over the world.

We offer young researchers an international and interdisciplinary setting for ambitious scientific projects, and we provide the appropriate training and working environment for a variety of technical and administrative professions.

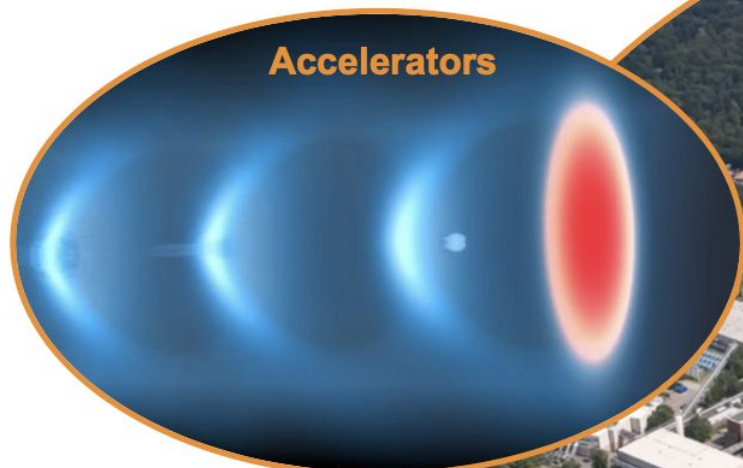


© DESY /  
R. Schaaf

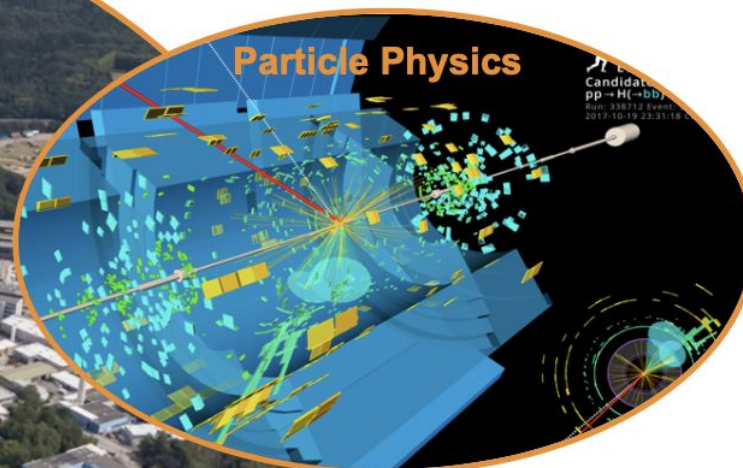


# DESY Activities

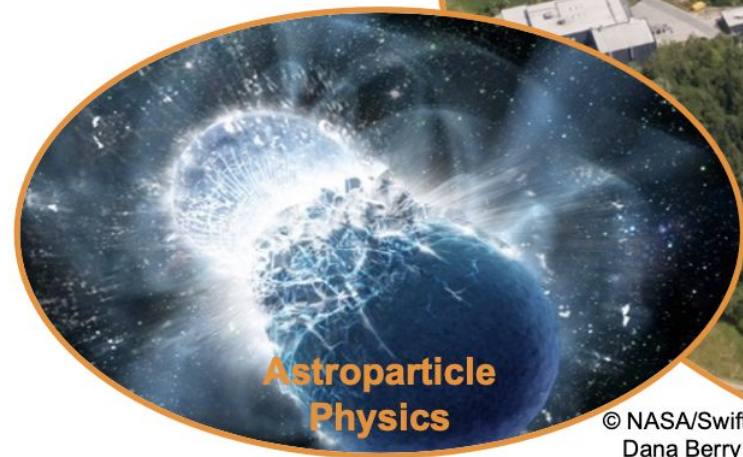
*From four research areas*



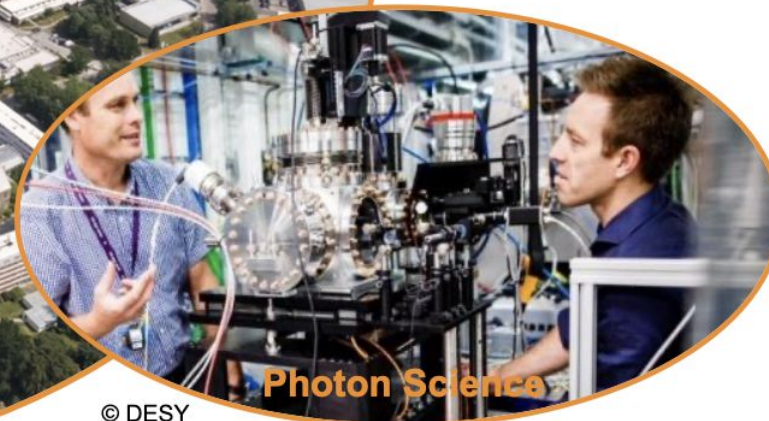
© DESY/UHH/A. Pousan



© ATLAS/CERN

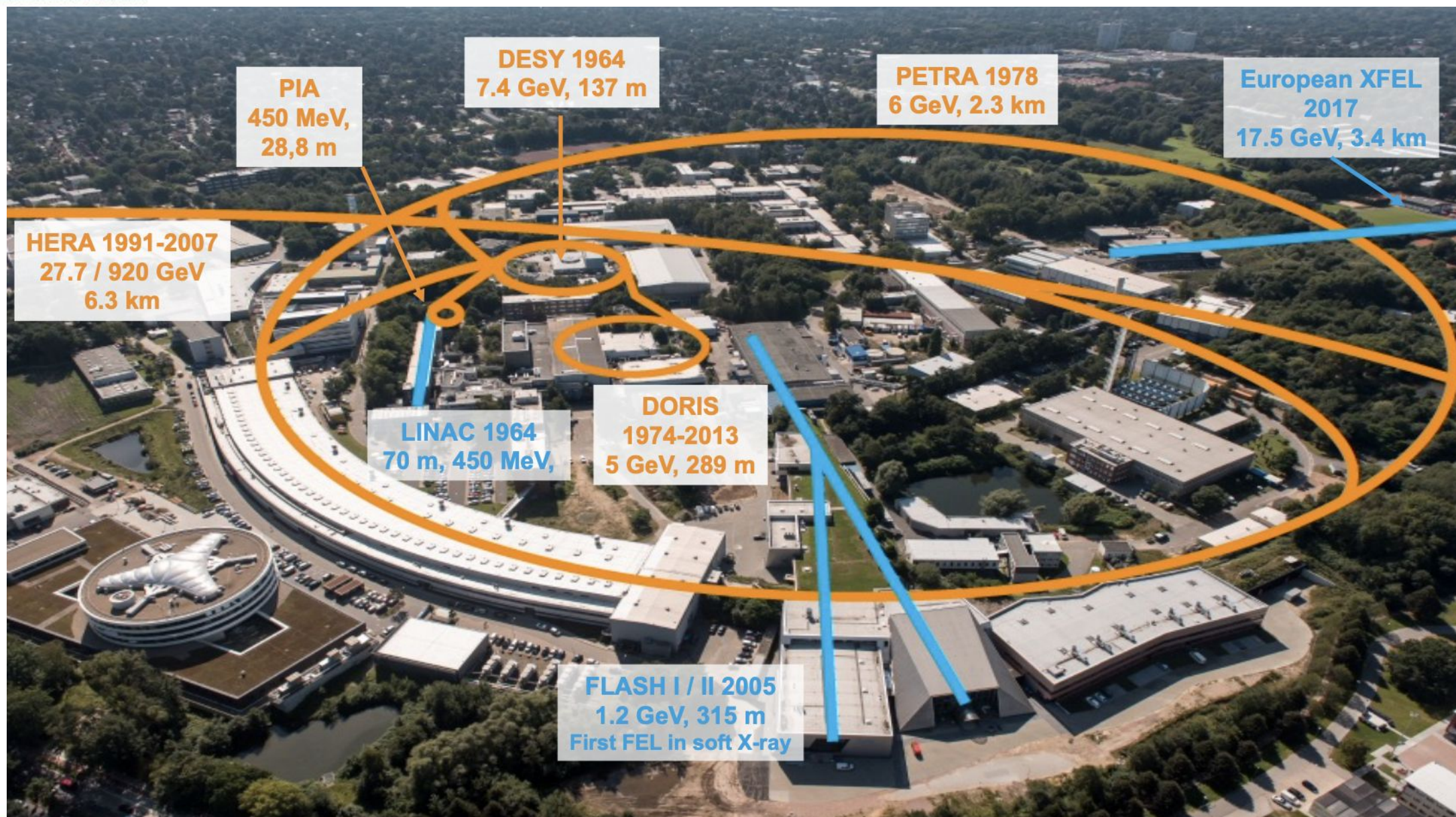


© NASA/Swift/  
Dana Berry



© DESY







# DESY today

## Interdisciplinary Research Campus

### National



### International





# DESY in Future

The vision for tomorrow: Science City Bahrenfeld



- Upgrade Petra III → Petra IV; second XFEL fan
- 123 ha development area, 13 research institutions (DESY, UHH, MPSD, FhG, UKE, HPI, BNITM, EMBL, FZB, HZI, FZJ, MHH, HZG)
- 9,500 researchers, technicians and administrative employees will work at site, > 3000 guests / year
- 4,000 students from physics, chemistry, biology
- **Innovative Ecosystem:** industry beamlines, innovation / technology centres, space for Entrepreneurs



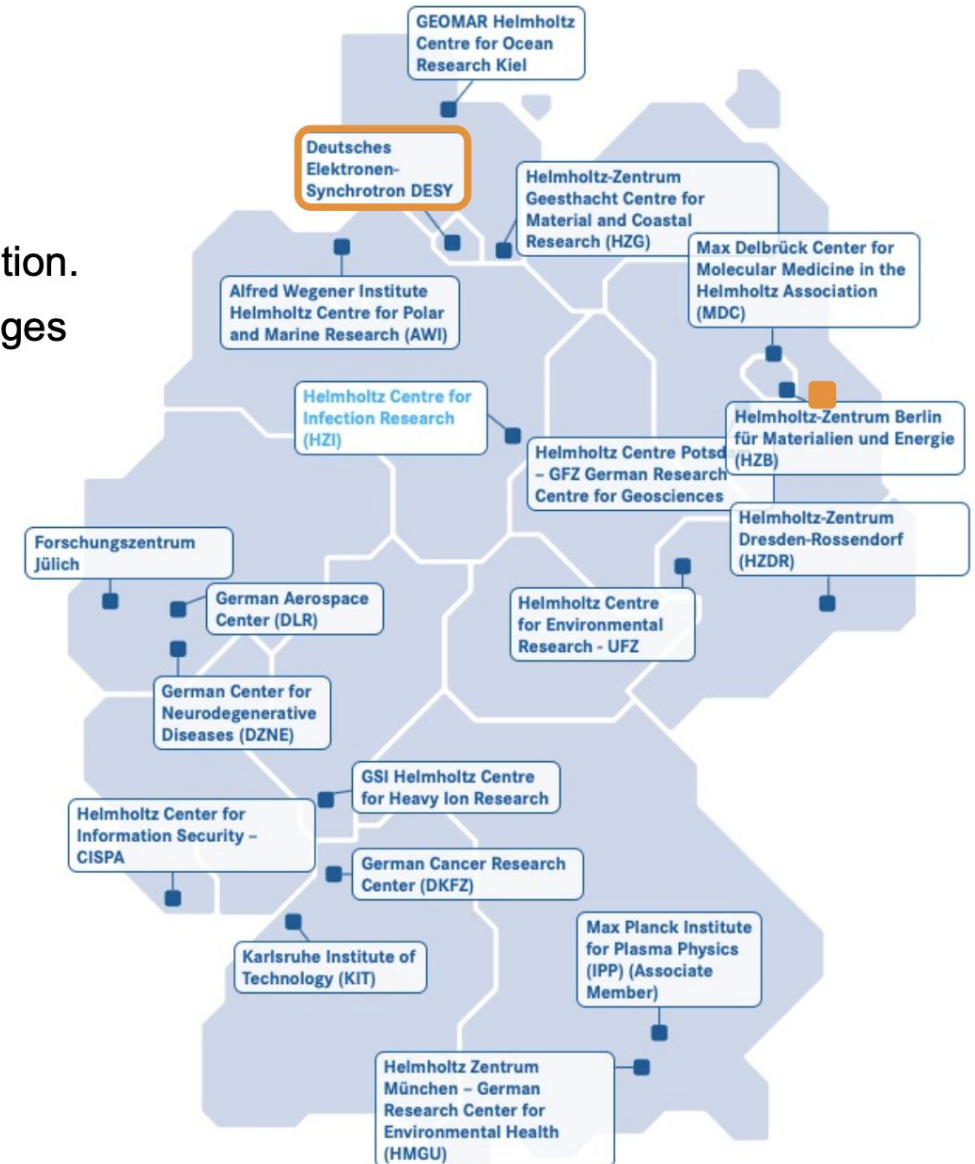
# Helmholtz and DESY

## The Helmholtz Association

**Mission:** Pursue the long-term research goals of state and society to maintain and improve the livelihoods of the population.

**Top-level research** to identify and explore the major challenges facing society, science and the economy.

**In numbers:** 6 research fields, 19 centres, > 40.000 staff, ~ 4.5 billion Euro budget → largest science organisation in Germany







ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

# Project overview / update

## *Pilot Phase Kick-Off*

João Fernandes (CERN)

Contact: [joao.fernandes@cern.ch](mailto:joao.fernandes@cern.ch)



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

**Starting Date:** 1<sup>st</sup> of January 2019

**Duration:** 42 Months

**Coordinator:** CERN (Lead Procurer)



European Commission



EMBL-EBI



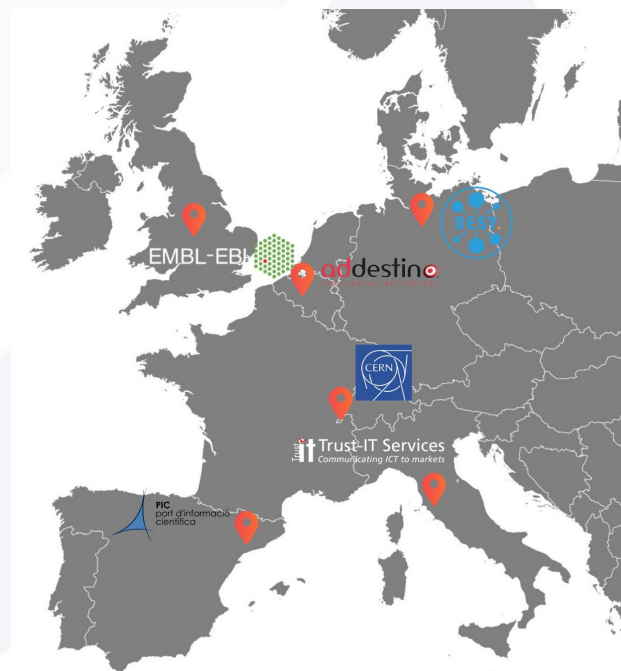
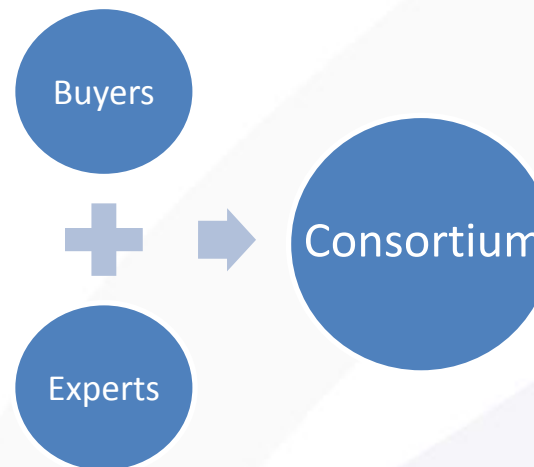
PIC  
port d'informació  
científica

**Buyers Group (BG)** - Public organisations committing funds to contribute to a joint-R&D-procurement, research data use cases and R&D testing effort

addestino  
innovation delivered.



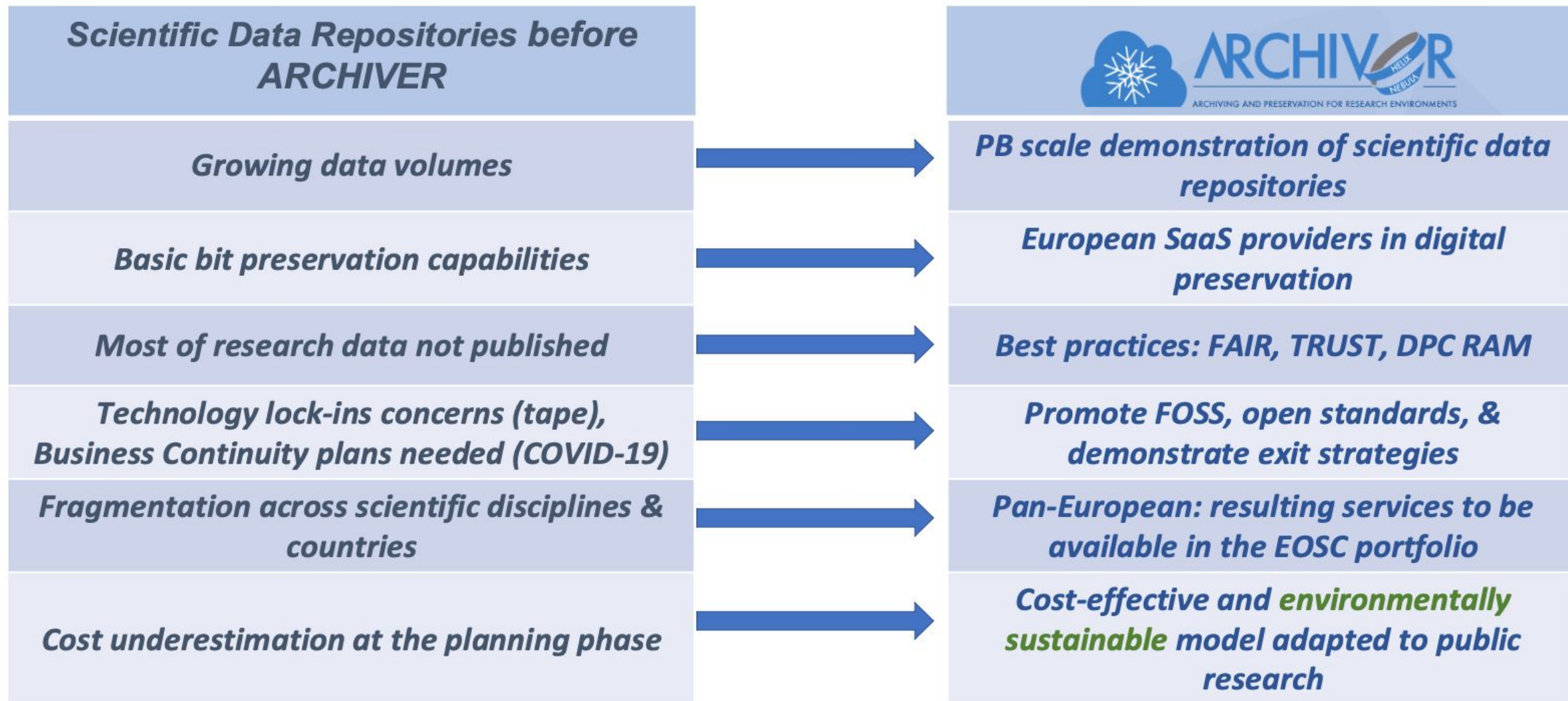
**Experts** - Partner organisations bringing expertise in requirement assessment and promotion activities



ARCHIVER has received funding from the European Union's H2020 Research & Innovation programme under Grant Agreement No 824516.

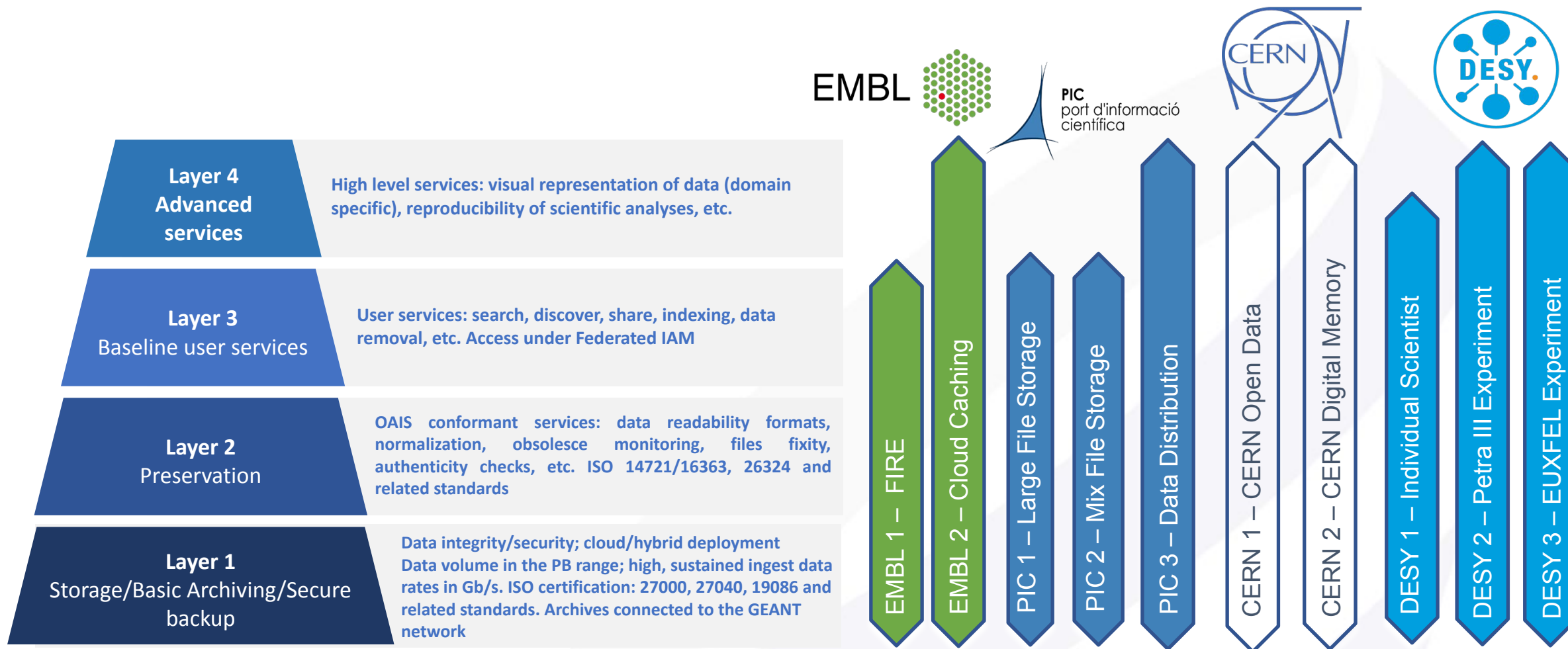


# Progress Beyond the State of the Art



ARCHIVER “current state of the art” report: <https://doi.org/10.5281/zenodo.3618215>

# R&D Scope

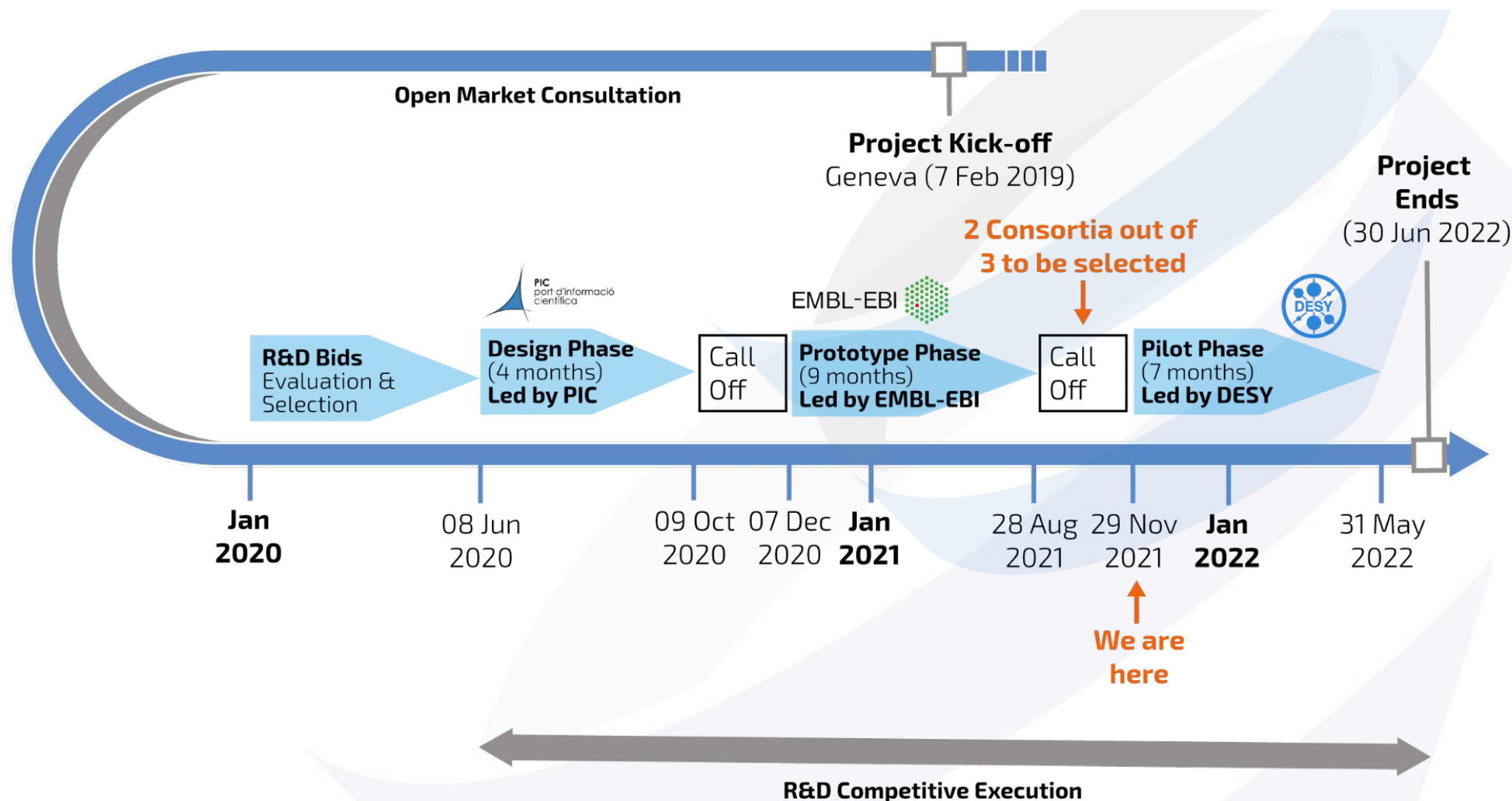


Scientific use cases deployments: <https://www.archiver-project.eu/deployment-scenarios>

ARCHIVER “current state of the art” report in the context of the EOSC: <https://doi.org/10.5281/zenodo.3618215>



# ARCHIVER Timeline



# Prototype Phase Highlights



- Functional objectives of the Prototype phase were successfully met.
- R&D challenge and scientific use cases requirements were globally understood.
- Systematic and structured feedback across the Buyers and Contractors to allow full understanding of the R&D challenge by the Contractors and its service capabilities by the Buyer organizations.
- Prototype Phase organized into 13 bi-weekly sprints in an Agile-like process with a regular rhythm of feedback and adjustment between Buyers and Contractors.
- CERN, EMBL-EBI, DESY & PIC allocated significant effort assessing and testing the prototypes, ingesting several hundreds of TBs of data.

# From Prototypes to Pilot



- R&D produced by the Contractors
  - More than 200 tests executed during the Prototype phase across three platforms
- Technical assessment in two steps:
  - R&D produced during the Prototype completed & of sufficient quality
  - Mini-competition (Call-off) for Pilot phase services R&D
- Main Criteria considered in the R&D review process:
  - Performance to scale correctly in the PB data region
  - R&D validation progressing from functional to “Go-To-Market” ready
  - Expertise in supporting Data Stewards achieving certification of scientific repositories
  - Clear commercialisation plans for the resulting services after the project end

# Pilot Phase Challenges



- Deployment model: on-premise vs cloud, portability, migration between different public cloud infrastructure, strategies to avoid vendor lock-ins, disaster recovery
- Advanced AAI & Access management; Network performance: 100TB/day
- Scalability and Elasticity: High Volumes for data ingestion and data recall
- Security: simulated cyberattacks
- Layer-4 reproducibility of scientific analyses independent of infrastructure
- FAIRness measurements: F-UJI tool: <https://www.f-uji.net/>
- Total Cost of Ownership (TCO): SaaS on cloud vs SaaS on-premise



**FAIRSFair**  
Fostering Fair Data Practices in Europe



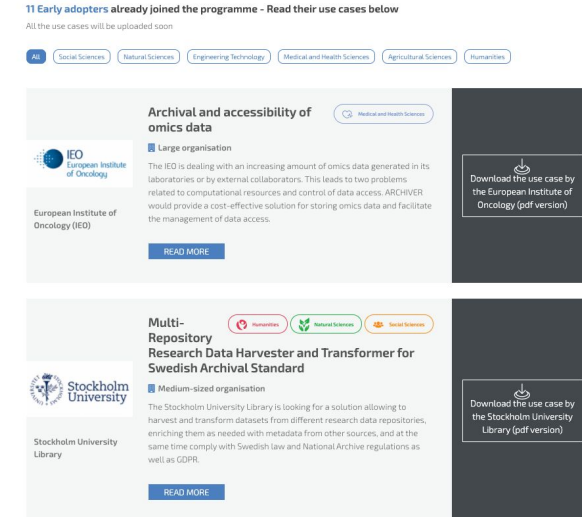
# Early Adopters <https://archiver-project.eu/early-adopters-programme>

- **Participants:**

- Demand side public sector organisations

- **Key advantages**

- Assess if resulting services address their archiving and preservation needs
- Contribute to shaping the R&D carried out in the project, identify additional use cases
- Have the option to purchase pilot-scale services by the end of the project



Friedrich Miescher Institute  
for Biomedical Research



Science and  
Technology  
Facilities Council



# A Minimum Viable EOSC

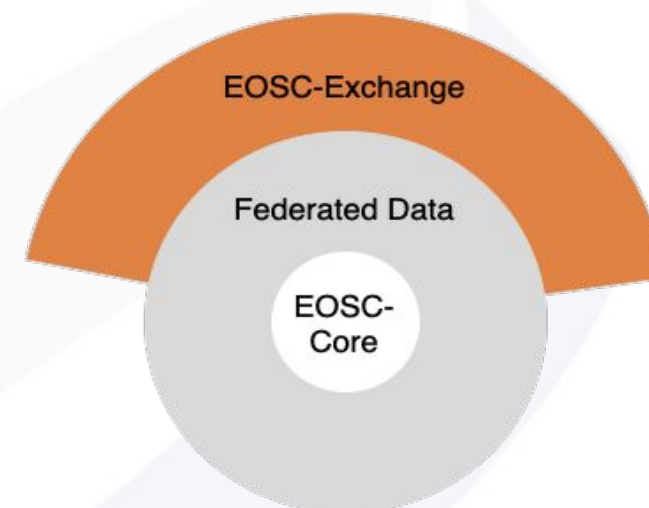
- EOSC Association established to govern the European Open Science Cloud
- Ambition to develop “Web of FAIR Data and Services” for science in Europe for multiple domains
- Provide access to services for storage, computation, analysis, *preservation*, etc. based on standards, combining data and services

“FAIR Forever Report: A study to remember” with an explicit mention to the ARCHIVER activities :

→ <https://www.dpconline.org/blog/fair-forever-a-fair-study-to-remember>

**EOSC Long-term Data Preservation Task Force charter**

→ [https://www.eosc.eu/sites/default/files/tfcharters/eosca\\_tflongtermdatapreservation\\_draftcharter\\_20210614.pdf](https://www.eosc.eu/sites/default/files/tfcharters/eosca_tflongtermdatapreservation_draftcharter_20210614.pdf)



## FAIR Forever?

Long Term Data Preservation Roles and Responsibilities, Final Report

February 2021 (V.7)

Dr Amy Currie and Dr William Kilbride

 EOSC FAIR Forever has received funding from the European Union under the EOSC Secretariat project. EOSCsecretariat.eu has received funding from the European Union's Horizon Programme call H2020-INFRAEOSC-01-2018-2019, grant Agreement number 831644. Europeana is an initiative of the European Union, financed by the European Union's Connecting Europe Facility and European Union Member States (<https://europeana.eu>) and <https://www.europeana.eu>

Charter for the EOSC - Task Force - Long Term Data Preservation (EOSC TF LTP)  
Version 0.5 (08-06-2021)

**Main aims**  
The possibility to reproduce, replicate and re-use scientific results depends on the long-term accessibility and assessability of the underlying data. The Strategic Research and Innovation Agenda (SRIA) of the EOSC underlines the importance of long-term data preservation, but an explicit strategy has not been formulated.

The EOSC TF LTP will provide recommendations for the EOSC board on the vision and sustainable implementation of long-term data preservation policies and practices, as well as suggestions to later strategy execution. It will address the roles and responsibilities of the different stakeholders, the financial aspects of long-term preservation and the necessary service infrastructure.

In this charter, long-term preservation is defined as a process for continued access to digital materials, or at least to the information contained in them, indefinitely.

**Core activities**  
1) **Creation of a shared understanding and vision**, starting from what we mean by digital preservation in the context of EOSC and a mapping of the existing landscape, resulting in a suggestion for a strategy, where horizontal EOSC preservation policy enables the connection and collaboration on national, community and local level. This TF will provide recommendations for the EOSC board on the vision and sustainable implementation of long-term data preservation policies and practices, as well as suggestions to later strategy execution.  
2) **Mapping and promotion of the roles, responsibilities and accountability** of the actors within the EOSC ecosystem with respect to long-term data preservation and responsibility levels defined in the SRIA. Complementing this by identifying the stakeholders in the different stages of the research data life cycle and their respective roles and responsibilities, including recommendations for awareness raising and training, especially on comprehensive and accountable DMPs.  
3) **Mapping of the financial aspects of long-term data preservation**, including the cost of digital preservation, financial responsibilities of the different stakeholders, as well as possible business models for repositories.  
4) **Recommendations on the creation of a European network of trustworthy digital repositories** following FAIR-enabling principles with disciplinary and geographical spread. Recommendations for EOSC data services to connect to this network and a roadmap to further mature the LTP aspects of these repository services.

**Working methodology**

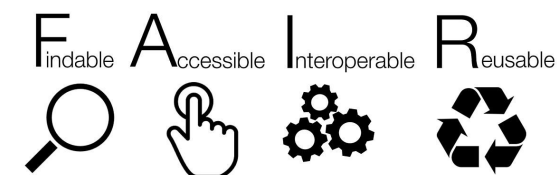
<sup>1</sup> <https://www.dpconline.org/handbook/lossy>

- Broad pan-European requirement analysis of the research sector
  - Analysis considered in the competitive R&D tender
  - Technical and organisational measures aligned with European legislation in the services being developed (by default & by design)
- Early Adopters Programme established
  - Additional use cases expanding further the set of supported scientific domains
  - Publicly funded research actors external to the ARCHIVER consortium
- Model for facilitate procurement of sustainable pilot services
  - Consortium members and Early Adopter organisations
  - Beyond the lifetime of the project

*ARCHIVER is the only EOSC related H2020 project focusing on Archiving & LTDP services for PB scale datasets across multiple research domains and countries.*

# Summary

- The R&D challenge of digital archiving goes **beyond data storage**: keep intellectual control of data and associated products for decades, make research outputs reusable
- Extending **FAIR** to research associated products: software, workflows, services and even infrastructures
- Acting as a template to **commoditise** archiving and preservation at scale in research domains with expert European SMEs
- ARCHIVER is promoting a **sustainable model** with services that will exist beyond the project lifetime in the context of the **EOSC**
- ARCHIVER pilot phase: validating services with end-users and early adopters organisations to determine suitability to their needs.





## Event Outline

14:30 - 14:40: Welcome from Sergey Yakubov (DESY)

14:40 - 15:00: Project overview / update - João Fernandes (CERN)

15:00 - 15:30: Expected outcomes of the Pilot Phase - Buyers Group representatives (CERN, DESY, EMBL-EBI, PIC)

15:30 - 15:40: Interactive brainstorming

*15:40 - 15:50: Break*



## Award ceremony

15:50 - 16:20: Presentation from consortium 1

16:20 - 16:50: Presentation from consortium 2

16:50 - 17:00: Closing remarks - João Fernandes (CERN)





ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

# Thank you!



[info@archiver-project.eu](mailto:info@archiver-project.eu)



<https://www.archiver-project.eu/>



<https://twitter.com/ArchiverProject>



<https://www.linkedin.com/company/archiver-project/>



<https://www.youtube.com/channel/UCCBIyLpUt-hWmQatqdlhIzw>



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.





ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

# Expected outcomes of the Pilot Phase

Buyers Group representatives  
CERN, DESY, EMBL-EBI, PIC



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.

# CERN Requirements and Expectations

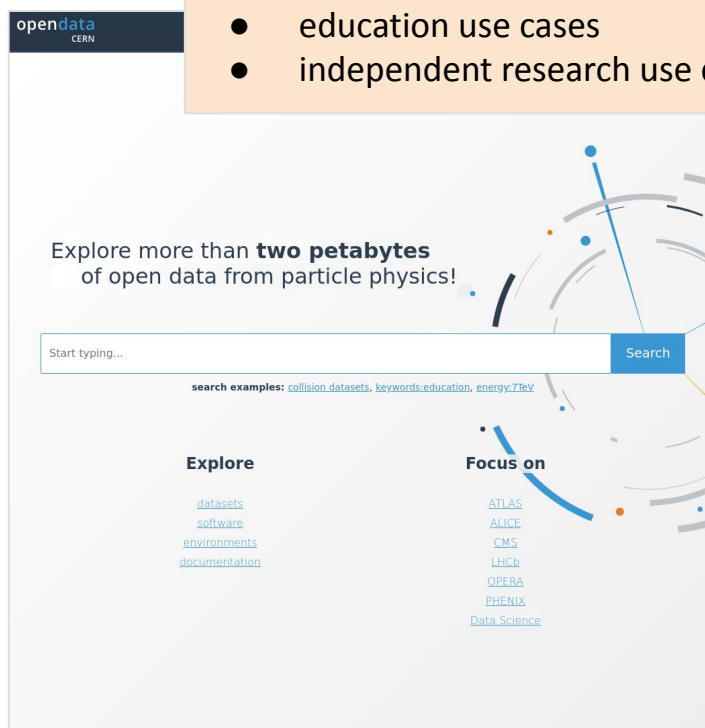
Tibor Simko, Jean-Yves Le Meur, Antonio Vivace, Ignacio Peluaga



# CERN Open Data: From data preservation to data reuse

## CERN Open Data portal

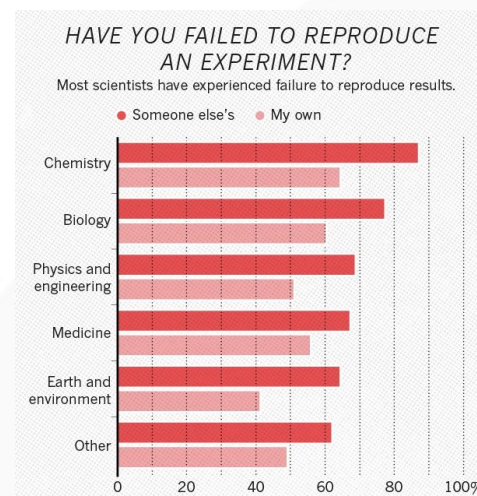
- more than 2.5 petabytes of particle physics data
- education use cases
- independent research use cases



<https://opendata.cern.ch>

## REANA reproducible analysis platform

- run containerised computational workflows on remote clouds
- CWL, Snakemake, Yadage
- HTCondor, Kubernetes, Slurm



Nature **533** (2016), 452–454

<https://doi.org/10.1038/533452a>

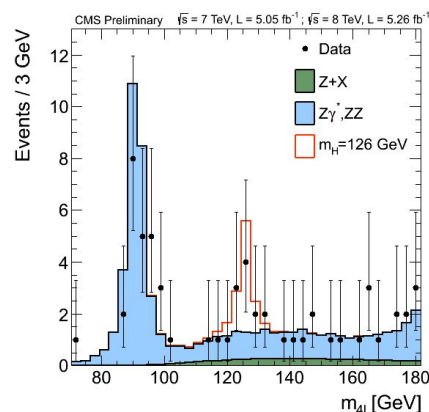
**Reproducible research data analysis platform**

<b>Flexible</b> Run many computational workflow engines.	<b>Scalable</b> Support for remote compute clouds.	<b>Reusable</b> Containerise once, reuse elsewhere. Cloud-native.	<b>Free</b> Free Software. MIT licence. Made with ❤️ at CERN.

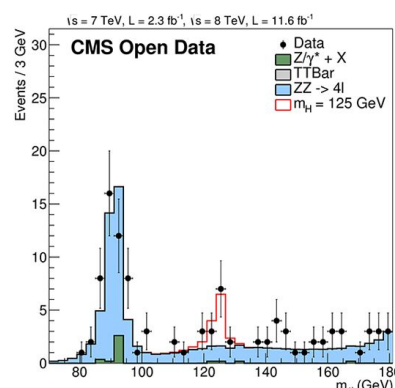
<https://www.reana.io>

# CERN Open Data: From data preservation to data reuse

- looking at “Compute” possibilities for content preserved in “Storage”
- reproducing several open data analysis examples
- reprocessing published simulated datasets
- challenges:
  - use containerised workflows (Docker, Kubernetes)
  - serve database content (CVMFS)

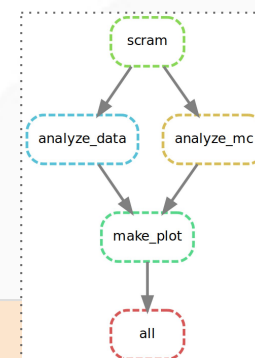


CMS Higgs-to-four-lepton example analysis



reana

```
rule analyze_data:
    input:
        config["data"],
        config["code"],
        "results/scramdone.txt"
    output:
        "results/DoubleMuParked2012C_10000_Higgs.root"
    container:
        "docker://cmsopendata/cmssw_5_3_32"
    shell:
        "source /opt/cms/cmssw_5_3_32/default.sh"
        "&& cd CMSSW_5_3_32/src"
        "&& eval `scramv1 runtime -sh`"
        "&& cd HiggsExample20112012/HiggsDemoAnalyzer"
        "&& cd ../Level3"
        "&& cmsRun demoanalyzer_cfg_level3data.py"
```



Running containerised workflows

Analysis of Higgs boson decays to two tau leptons using data and simulation of events at the CMS detector from 2012

Wunsch, Stefan

Cite as: Wunsch, Stefan; (2019). Analysis of Higgs boson decays to two tau leptons using data and simulation of events at the CMS detector from 2012. CERN Open Data Portal. DOI:10.7483/OPENDATA.CMS.GV20.PR5T

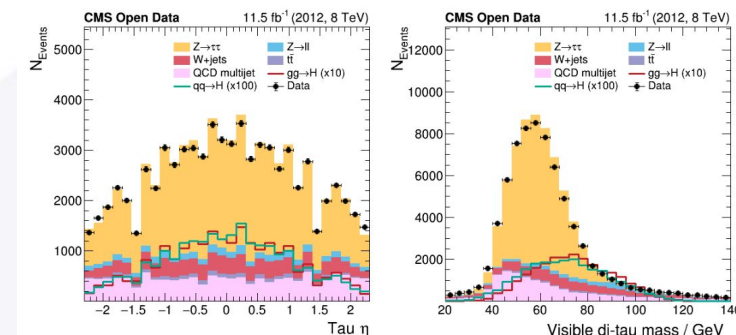
Software Analysis Workflow CMS CERN-LHC

## Description

This analysis uses data and simulation of events at the CMS experiment from 2012 with the goal to study decays of a Higgs boson into two tau leptons in the final state of a muon lepton and a hadronically decayed tau lepton. The analysis follows loosely the setup of the official CMS analysis published in 2014.

The purpose of the original CMS analysis was to establish the existence of the Higgs boson decaying into two tau leptons. Since performing this analysis properly with full consideration of all systematic uncertainties is an enormously complex task, we reduce this analysis to the qualitative study of the kinematics and properties of such events without a statistical analysis. However, as you can explore in this record, already such a reduced analysis is complex and requires extensive physics knowledge, which makes this a perfect first look into the procedures required to claim the evidence or existence of a new particle.

Two example results produced by this analysis can be seen below. The plots show the data recorded by the detector compared to the estimation of the contributing processes, which are explained in the following. The analysis has implemented the visualization of 34 such observables.

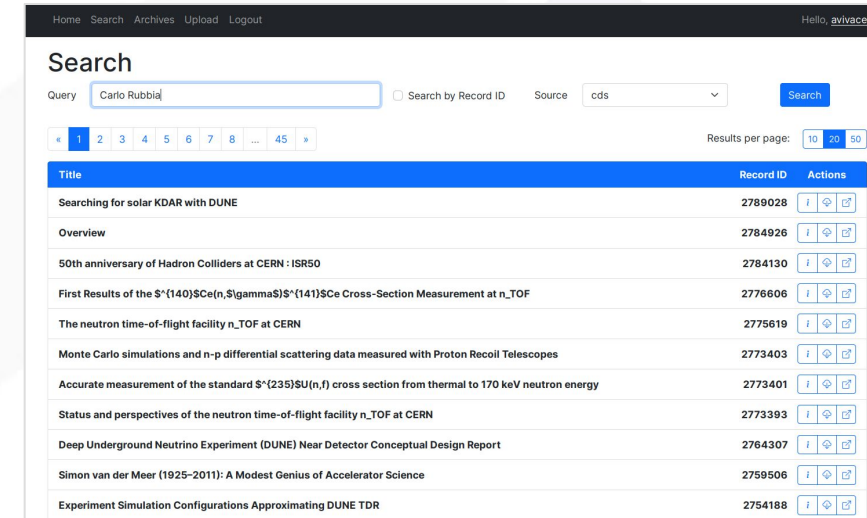


CMS Higgs-to-tautau example analysis



# CERN Digital Memory: the OAIS Archive

- **Status:**
  - Tool to harvest main CERN Information Systems and create SIP (following BagIt specs).
    - Includes publications, preprints, presentations, photos, videos, and more
  - Control Interface to run “on-demand” archiving for people leaving CERN
- **Challenges:**
  - Duplication
  - Authorships
  - Integrity
  - Versioning
  - Running preservation services/ file conversions to create AIPs



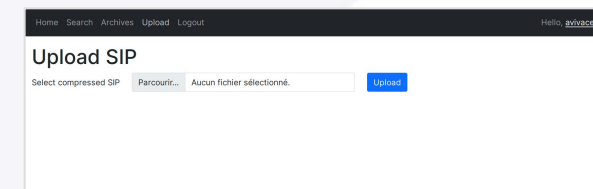
Home Search Archives Upload Logout Hello, avivace

### Search

Query:  ☐ Search by Record ID Source:

« 1 2 3 4 5 6 7 8 ... 45 » Results per page:  20 50

Title	Record ID	Actions
Searching for solar KDAR with DUNE	2789028	<a href="#">i</a> <a href="#">d</a> <a href="#">u</a>
Overview	2784926	<a href="#">i</a> <a href="#">d</a> <a href="#">u</a>
50th anniversary of Hadron Colliders at CERN : ISR50	2784130	<a href="#">i</a> <a href="#">d</a> <a href="#">u</a>
First Results of the $^{140}\text{Ce}(n,\gamma)^{141}\text{Ce}$ Cross-Section Measurement at n_TOF	2776606	<a href="#">i</a> <a href="#">d</a> <a href="#">u</a>
The neutron time-of-flight facility n_TOF at CERN	2775619	<a href="#">i</a> <a href="#">d</a> <a href="#">u</a>
Monte Carlo simulations and n-p differential scattering data measured with Proton Recoil Telescopes	2773403	<a href="#">i</a> <a href="#">d</a> <a href="#">u</a>
Accurate measurement of the standard $^{235}\text{U}(n,f)$ cross section from thermal to 170 keV neutron energy	2773401	<a href="#">i</a> <a href="#">d</a> <a href="#">u</a>
Status and perspectives of the neutron time-of-flight facility n_TOF at CERN	2773393	<a href="#">i</a> <a href="#">d</a> <a href="#">u</a>
Deep Underground Neutrino Experiment (DUNE) Near Detector Conceptual Design Report	2764307	<a href="#">i</a> <a href="#">d</a> <a href="#">u</a>
Simon van der Meer (1925-2011): A Modest Genius of Accelerator Science	2759506	<a href="#">i</a> <a href="#">d</a> <a href="#">u</a>
Experiment Simulation Configurations Approximating DUNE TDR	2754188	<a href="#">i</a> <a href="#">d</a> <a href="#">u</a>

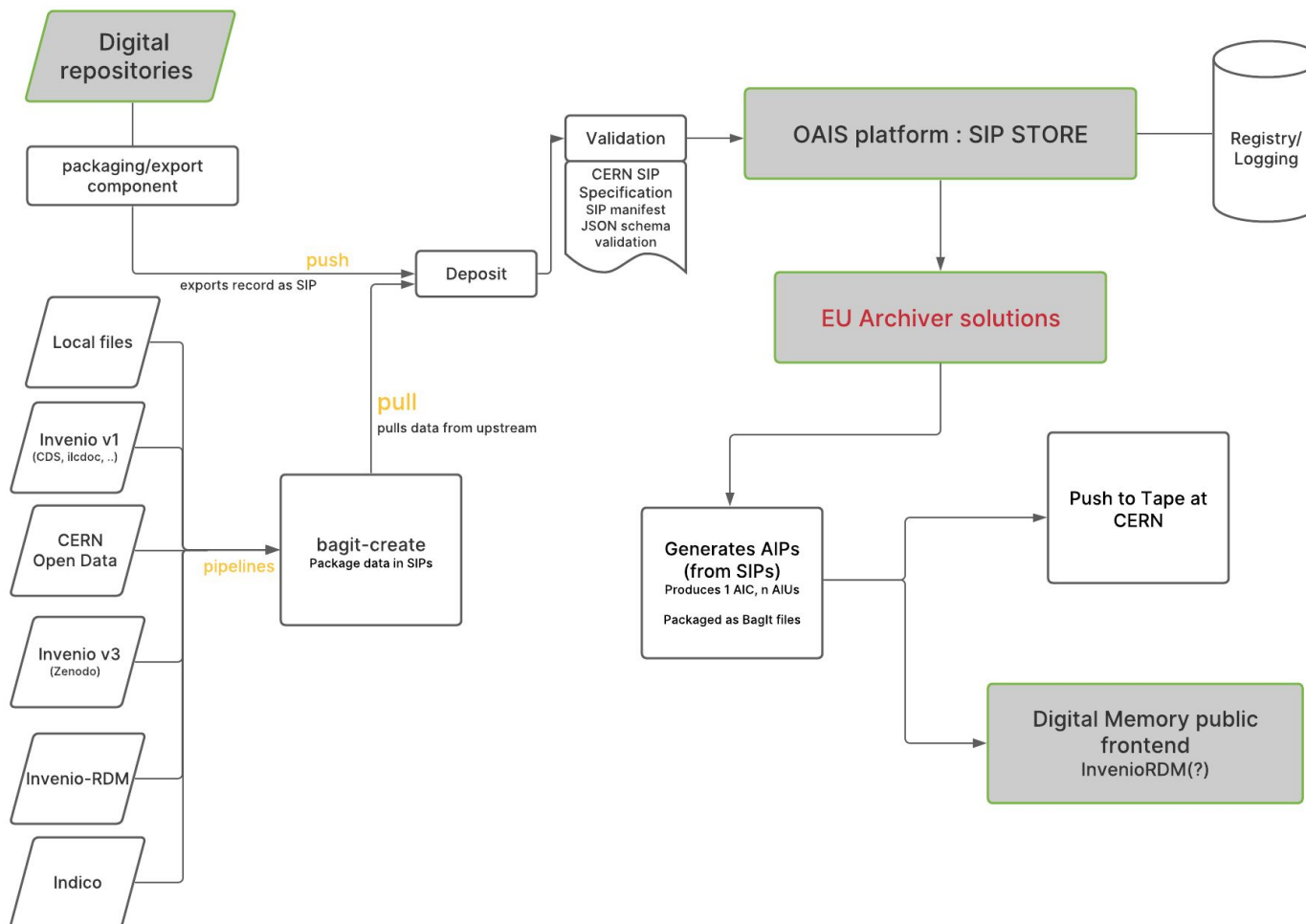


Home Search Archives Upload Logout Hello, avivace

### Upload SIP

Select compressed SIP  Aucun fichier sélectionné.

# CERN Digital Memory: feeding Archiver.eu



Archiver.eu solutions to ingest CERN SIPs and store artifacts (AIPs, DIPs...)

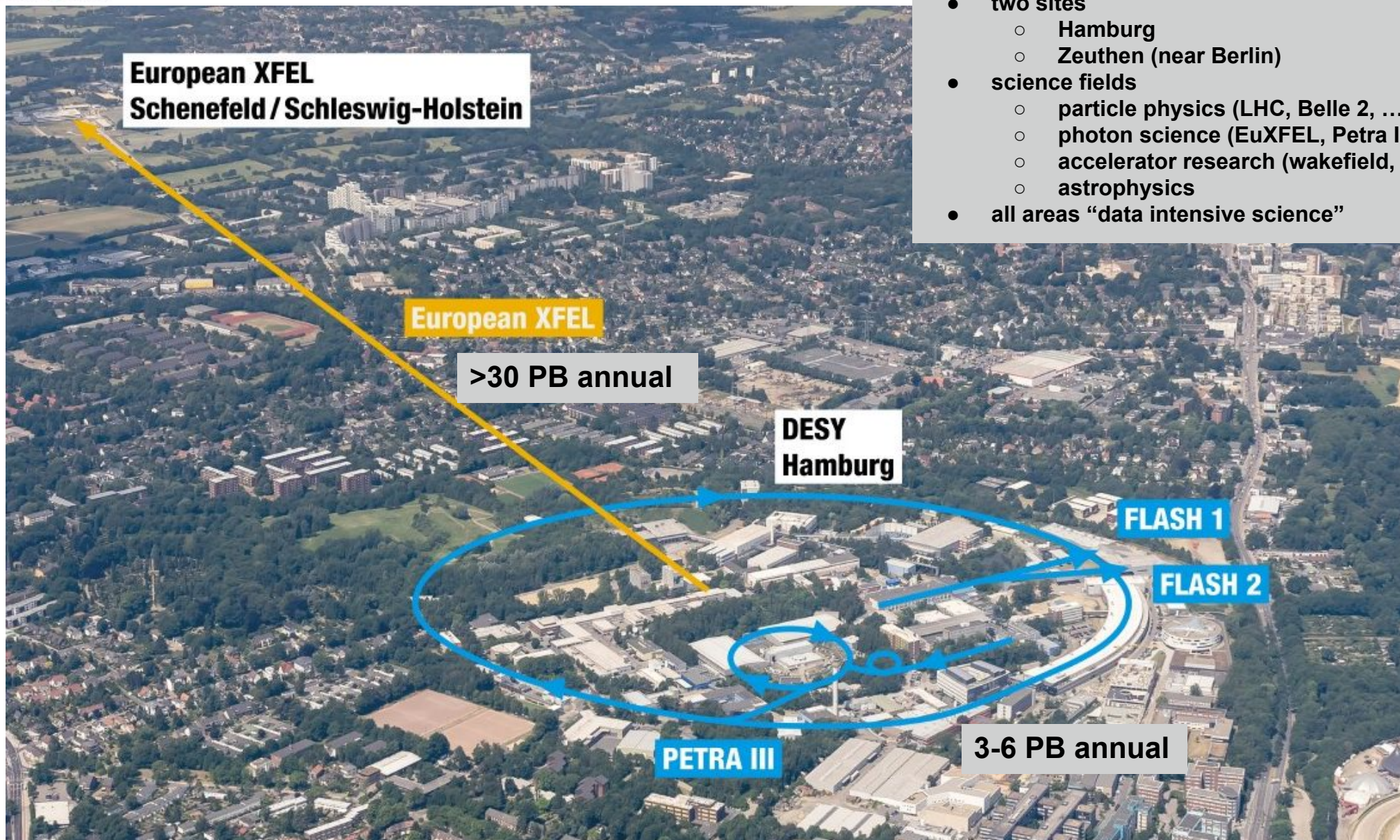


# DESY Requirements and Expectations

Sergey Yakubov, Martin Gasthuber



## Main sources of data to be archived and preserved - 2021

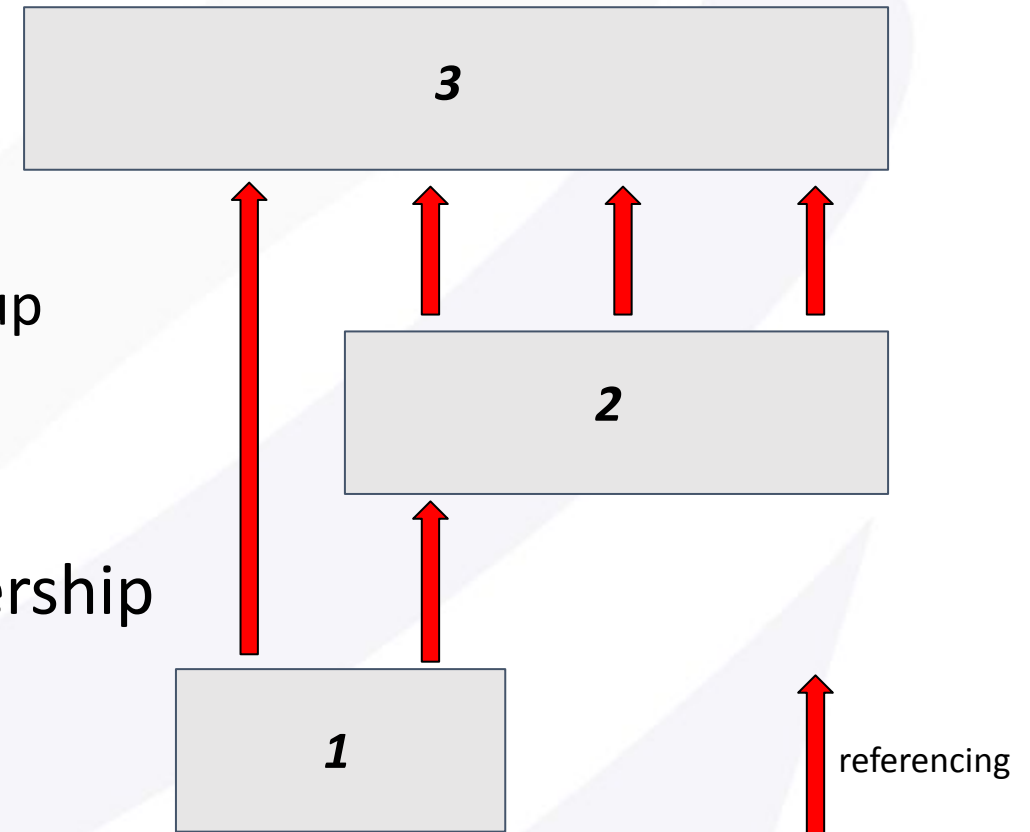


- two sites
  - Hamburg
  - Zeuthen (near Berlin)
- science fields
  - particle physics (LHC, Belle 2, ...)
  - photon science (EuXFEL, Petra III, FLASH)
  - accelerator research (wakefield, Petra IV, ...)
  - astrophysics
- all areas “data intensive science”



## use cases - selection and relations

- photon science focussed - pushed by 3 local independently running accelerator facilities
  - ground zero of the data deluge
- three, partly nested, scaled use cases
  - **(1)** individual scientist/small working group
  - **(2)** large working group/experiment
  - **(3)** facility level (with data policy)
- existing entities/actors with data ownership and stewardship responsibilities



## More tangible / general expectations

- core functionalities completed in last phase
- focus on scaling and stability validations
- Tape integration - scaling requires this 'low cost' storage technology
- hybrid deployments - i.e. MD and initial data copy on-site, secondary data and MD backup (cold & encrypted) off-site - targeting at:
  - off-site data co-located with public cloud computing for 'open data' and related analysis tasks
  - primary off-site data attractive to communities/groups not having an 'IT-Home'



## Case **1** - individual scientist / small working group

- decent volume & bandwidth requirements
  - few 10 TB, 100MB/s, 10K objects - per archive
  - ~0.2-0.5PB annual
- access mainly via a web browser (GUI)
  - pre-defined MD schema & policy and templates by admins
- scientist (in person) - owner and archivist
- challenges
  - authentication - federated AAI
  - metadata scheme - hierarchy of schemes, mandatory fields, controlled extension
  - DOI / publication
  - local & hybrid deployments
  - open-data management & access

## Case 2 - mid size - Petra III experiment

- Case **1** plus...
- increased volume and bandwidth requirements
  - ~100 TB, 1-2GB/s, >150K objects per archive
  - ~3-6 PB annual
- mainly API access (non interactive)
  - agent - automated life cycle process integration
- challenges
  - templating - i.e. MD, lifetimes, access rights, etc.
    - compliant to data policy
  - data storage costs starts to be important
    - volume and bandwidth *amplification* inside your solution - rising importance



## Case 3 - facility level - EuXFEL

- Case 2 plus...
- largest scale
  - few 100 TB up to 1-2 PB, 2-10GB/s, >30K objects - per experiment
    - 1.3PB per day raw data production seen recently - and growing
  - >30 PB annual
- non interactive - full API access
  - full templating for creation and filling process of the archive
- challenges
  - scale
  - data storage costs becomes the most important criteria
    - volume and bandwidth *amplification* inside your solution becomes dominating criteria

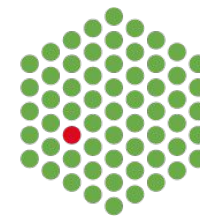
# EMBL-EBI Requirements and Expectations

Justin Clark-Casey

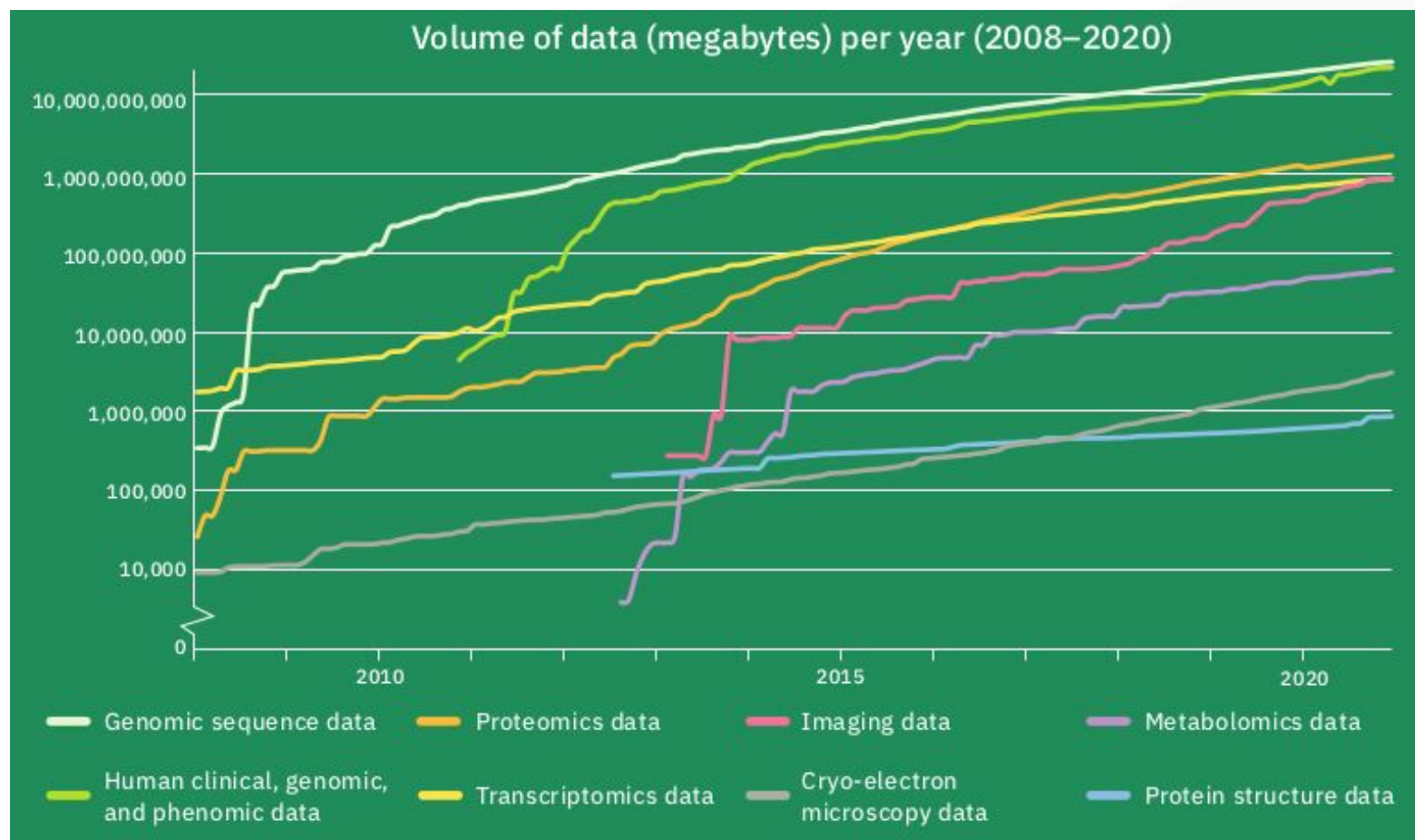


## About EMBL-EBI

- “The home for big data in biology”
- Datasets, datasets, datasets: genomics, proteomics, metagenomics, transcriptomics, imageomics ...
  - You (researchers) submit them
  - We curate them
  - We provide them
  - You download them/(cloud)-access them/analyze them
  - We archive them
- Part of the European Molecular Biology Laboratory (27 member states, 1800 ppl, 80 independent research groups)
  - An intergovernmental organization like CERN.

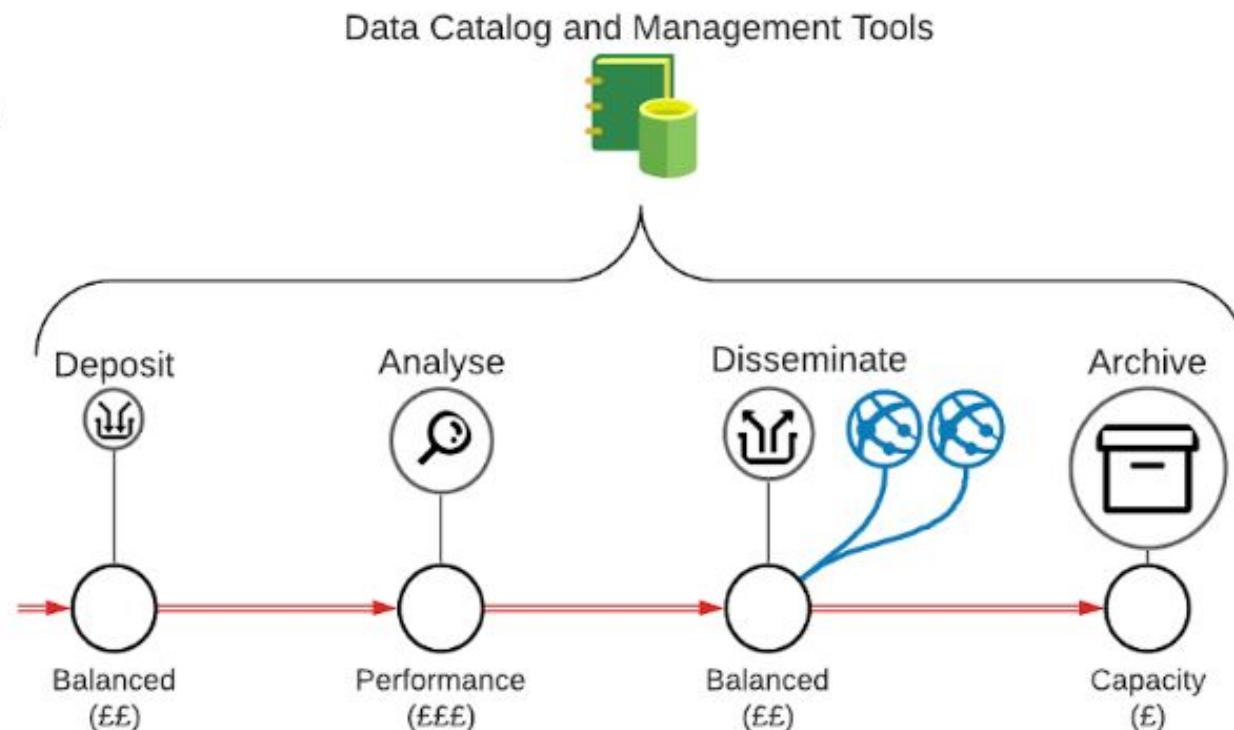


# Data growth at EMBL-EBI



## EMBL-EBI Archiver Pilot Phase Plan

- Pilot preservation of selected large datasets from across EMBL-EBI as part of an integrated data management system





## EMBL-EBI Archiver Pilot Phase Plan

- Deploy a pilot in a real storage environment
- Taking forward the EMBL on FIRE use case
- No new feature requests
  - But may use features developed for other buyers (e.g. flexible metadata schemas)
- Assess in pilot conditions
  - Ease of integration
  - Scalability
  - Security
  - Robustness
  - Cost

# PIC Requirements and Expectations

Jordi Casals, Manuel Delfino, Jordi Delgado

# Port d'Informació Científica

Use cases will be based on MAGIC Telescope data  
Observatorio del Roque de los Muchachos  
(La Palma, Canary Islands)

- Collecting data 365 days a year
  - Except when a volcano starts activity
- Hint: NOW
- 300TB per year for ranges of 5-6 years
- Random recalls during the period



\_\_\_\_\_ In collaboration with \_\_\_\_\_  
ALBA Synchrotron

- More than 10 Beamlines (and growing for next years)
- Datasets ranging from 200TB up to 4PB
- Internal and external scientific users





# Pilot Requirements

- Petabyte level Storage → functional, reliable, good performance, reasonable cost
  - From 1PB in 2021 to 15PB in 2025
  - Scalability testing → upload and downloads with production environments, sizes and bandwidth
- Integrate and automate processes using services CLI or API based scripts with production systems
- Create custom metadata schemas automatically on upload from data sources
  - Run production processes based on metadata searches
- Give access to users inside the collaboration with different permissions based on roles in experiment
- Test in-archive data processing using co-located Cloud
  - Enables automatic processing of uploaded data and prevents moving data around two times
- Test future reprocessing for reusability, based on custom containers, python notebooks or any possible system that may help the data to not expire
- Analyze production costs to evaluate if it would be a usable option

Everything will be tested by both PIC and ALBA and put in common to better evaluate the final results and possibilities of the resulting services

Any questions?

## **ARCHIVER PILOT PHASE AWARD**



**congratulating the winning consortia**

**Florian Pappenberger, Director of Forecasts**



Go to [www.menti.com](https://www.menti.com) and use the code  
3228 5110

# BREAK





ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

# PILOT PHASE AWARD CEREMONY

And the winners  
are.....



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.





arkivum

Bringing archived data to life



Google Cloud

libnova



UNIVERSITAT DE  
BARCELONA

voxility





ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

## Arkivum - Google Cloud



arkivum

Bringing archived data to life



Google Cloud



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

ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

## INTERVIEW WITH THE SELECTED BIDDERS



### arkivum

Bringing archived data to life



## Public Award Ceremony







arkivum

Bringing archived data to life

# SCALABLE AND SUSTAINABLE LONG TERM DIGITAL PRESERVATION OF SCIENTIFIC DATASETS

Matthew Addis

Arkivum

ARCHIVER Pilot Kick-Off Event 29 Nov 2021

# Contents

- Sustainability
  - Scalability, Performance and Cost Effectiveness
  - Portability, Open Standards, Data Sovereignty
  - FAIR Forever, Digital Preservation, Long-term Access and Reuse
  - Environmental Sustainability
- Summary

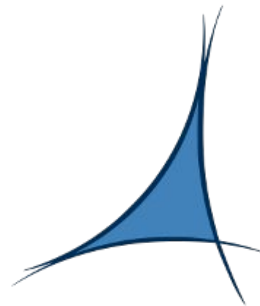
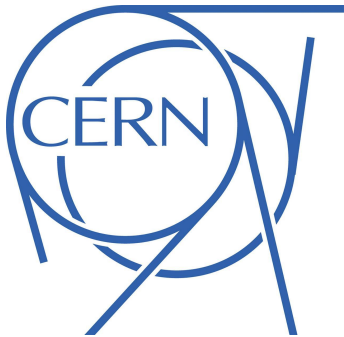
# Scalability, Performance and Cost Effectiveness





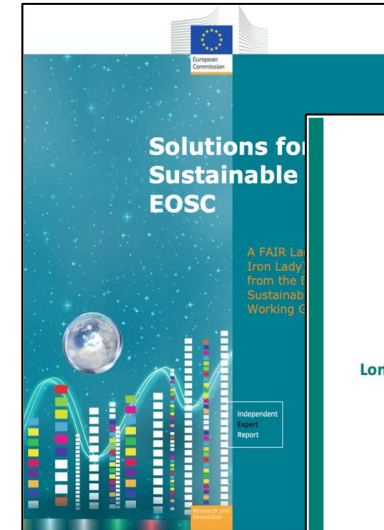
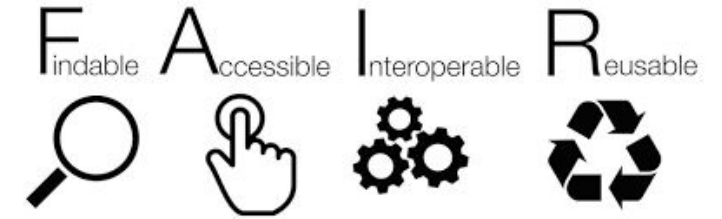
# ARCHIVER: requirements

- Archive size: 10-100 PB
- Ingest rates: 100TB per day
- Data types: raw data, derived data, metadata
- Users: people and software applications

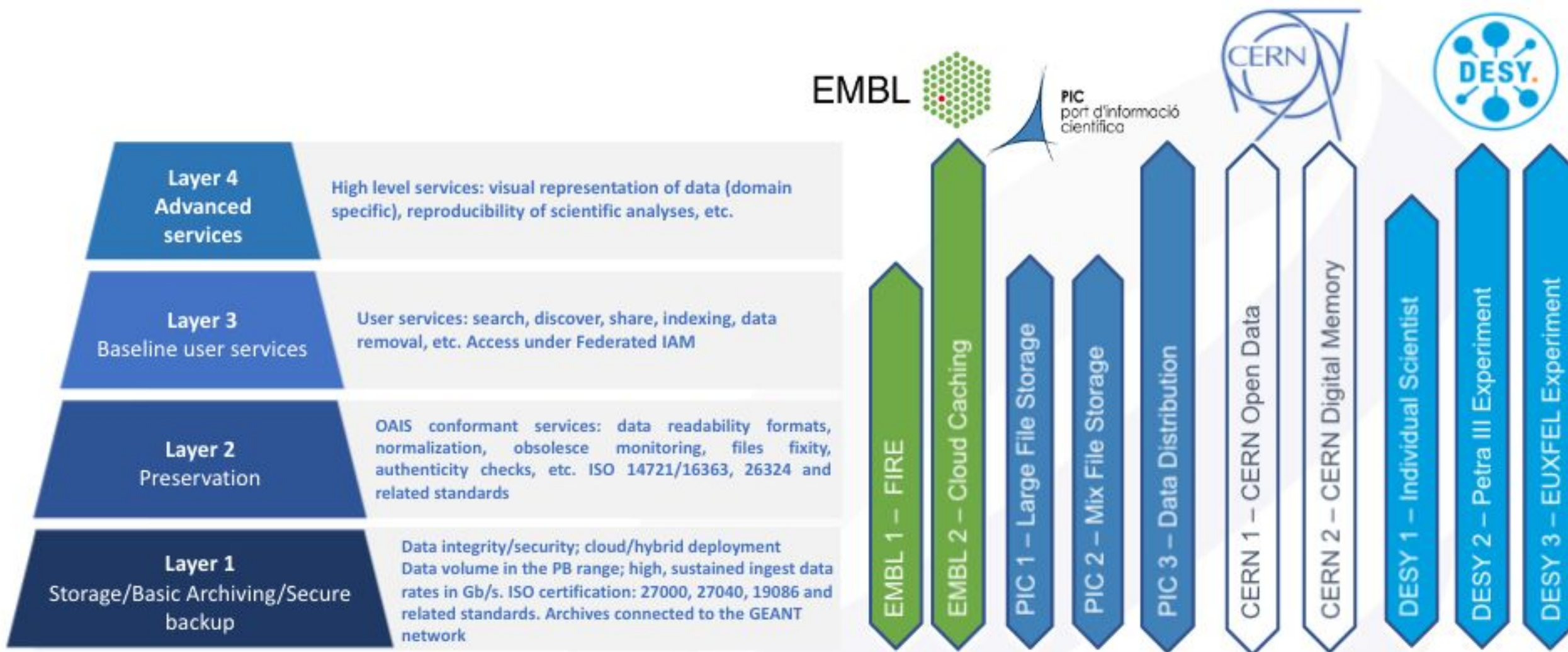


**PIC**  
port d'informació  
científica

EMBL-EBI



# Demand side requirements

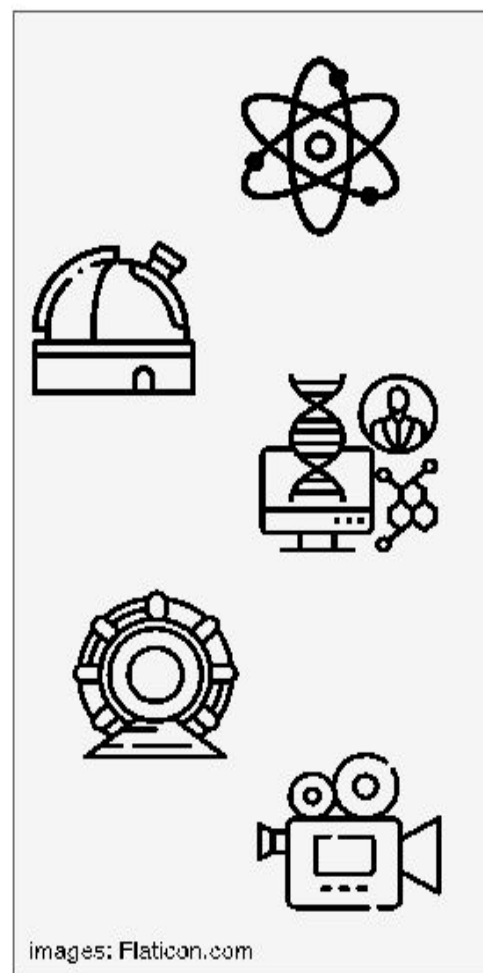


Scientific use cases deployments: <https://www.archiver-project.eu/deployment-scenarios>

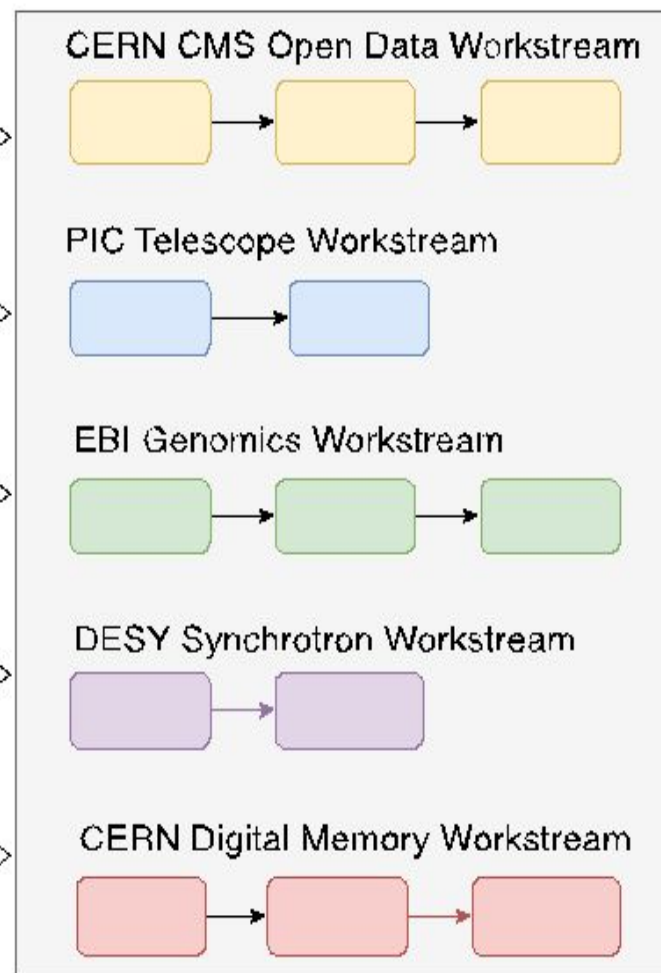
ARCHIVER "current state of the art" report in the context of the EOSC: <https://doi.org/10.5281/zenodo.3618215>

# Long-term digital preservation (LTDP) factories to support sustainable FAIR data

Content types and sources



Automated Workflows



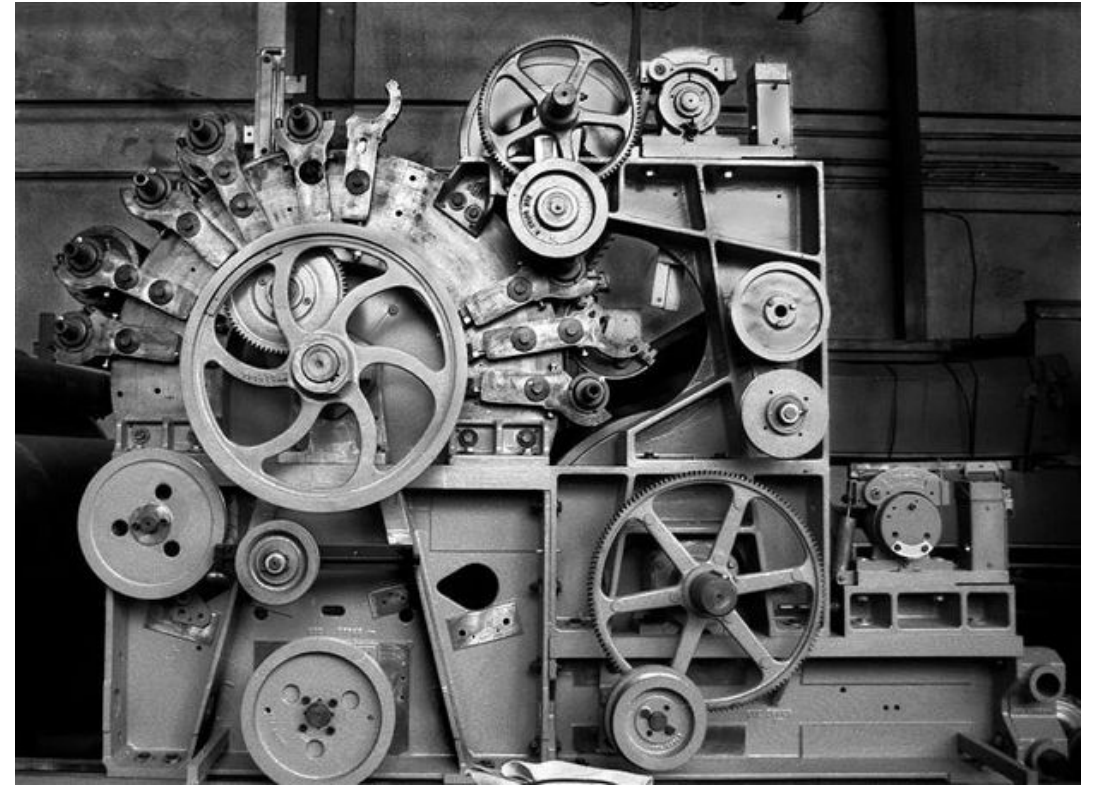
FAIR data for Researchers





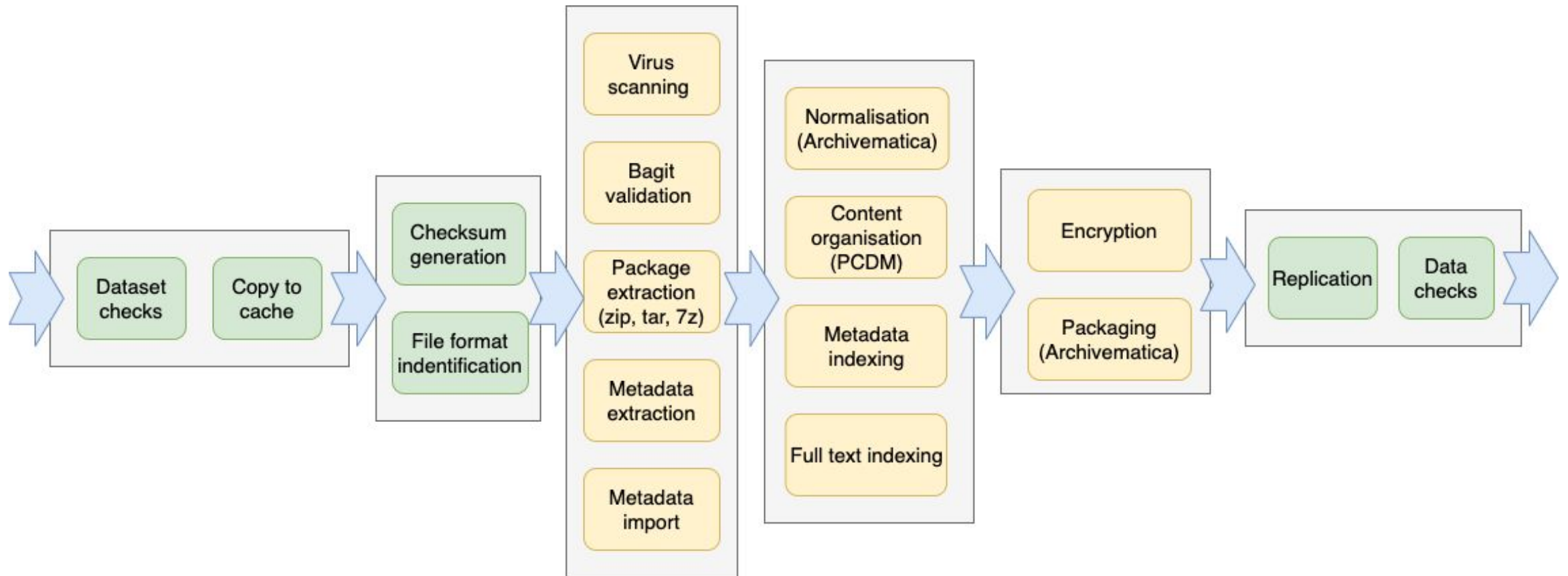
# LTDP factories that scale

- Automation
- Minimal Effort Ingest / Minimal Viable Preservation
- Scalable and efficient use of IT resources
  - Serverless computing
  - Microservices
  - Scale-out parallel processing
- Robust workflows
  - Failures will happen
  - Record and audit everything



Eugen Stoll, CC BY-NC 2.0, <https://flic.kr/p/5c8cz>

# Microservices, serverless computing, cloud infrastructure



Bringing archived data to life

# Ingest and archiving of a 50TB dataset

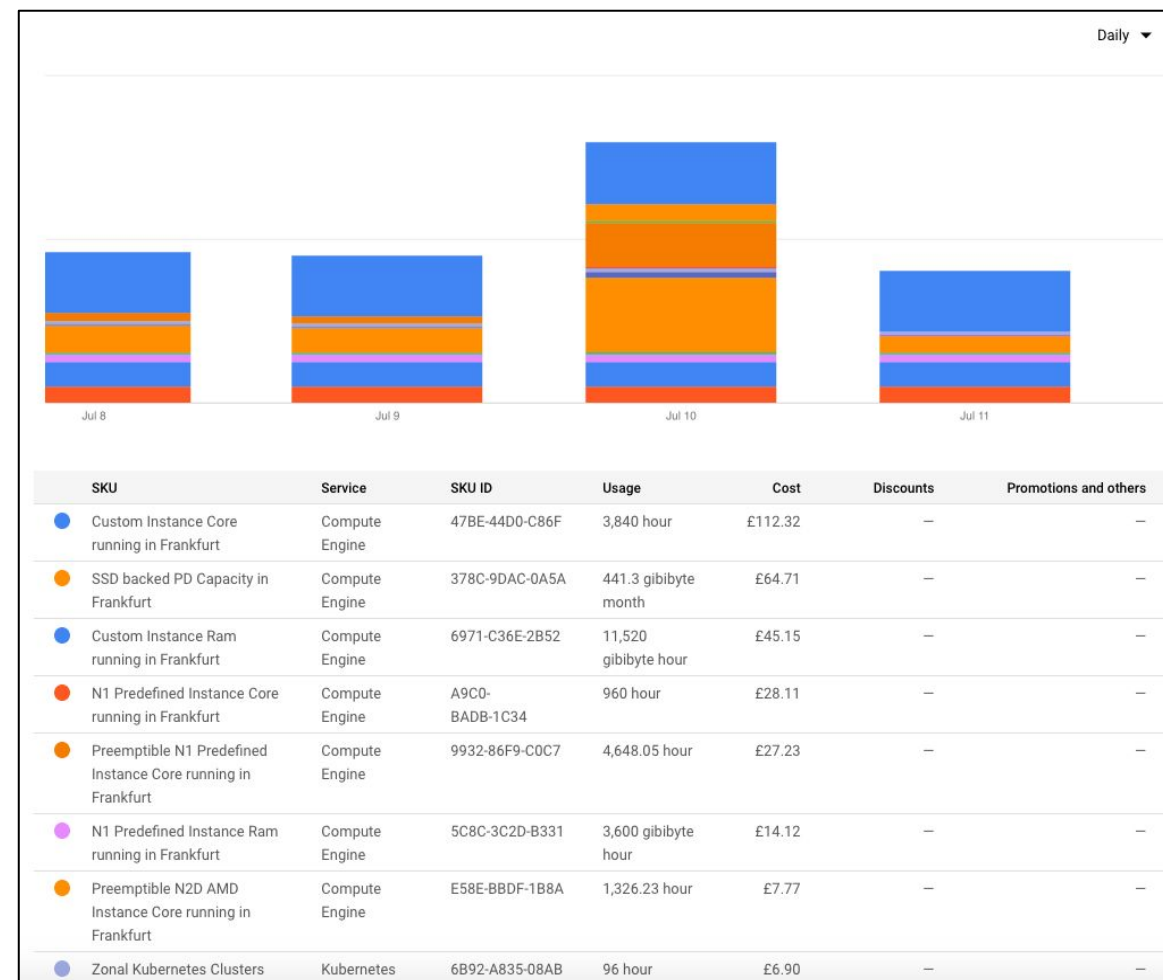


50,000 HDF5 files (50TB) ingested and stored at a rate of 150TB/day



# Costs and Benchmarking

- Execute real world scenarios on GCP
- Record parameters
  - execution time
  - data volumes, number of files
  - type of activity (ingest, export, preservation)
- Extract baseline costs from GCP
- Extract overhead of running scenario
- Normalise to create a 'cost per TB' metric
- Calculate short-term and long-term storage
  - Upload/export buckets, caching, archive buckets



# Costs Projections

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Ingest cost TB (\$)	3.27	Note: This will automatically include the extra storage needed if Archivematica is included in the ingest scenario
Storage cost TB/year (\$)	28.80	
Retrieval cost TB (\$)	50.00	
Export cost TB (\$)	1.30	

TB -> PB		
GB -> TB multiplier	1000	(Edit this on Conversion Factors Tab)

Year	Annual Data ingest (PB)	Cummulative Archived Data(PB)	Size band	Annual Export as % of Total Archived Data	Annual Exported Data (PB)	Annual Ingest Cost (\$)	Annual Storage Cost (\$)	Annual Retrieval Cost (\$)	Annual Export Cost (\$)	Total Annual Cost (\$)	Averaged Annual Cost per Archived TB (\$)
1	0.50	0.50	band1	5%	0.03	1635	14400.00	1250	32.50	17317.50	34.64
2	1.00	1.50	band2	5%	0.08	3270	43200.00	3750	97.50	50317.50	33.55
3	1.00	2.50	band2	5%	0.13	3270	72000.00	6250	162.50	81682.50	32.67
4	1.00	3.50	band2	5%	0.18	3270	100800.00	8750	227.50	113047.50	32.30
5	1.00	4.50	band2	5%	0.23	3270	129600.00	11250	292.50	144412.50	32.09
6	2.00	6.50	band2	5%	0.33	6540	187200.00	16250	422.50	210412.50	32.37
7	2.00	8.50	band2	5%	0.43	6540	244800.00	21250	552.50	273142.50	32.13
8	2.00	10.50	band3	5%	0.53	6540	302400.00	26250	682.50	335872.50	31.99
9	2.00	12.50	band3	5%	0.63	6540	360000.00	31250	812.50	398602.50	31.89
10	2.00	14.50	band3	5%	0.73	6540	417600.00	36250	942.50	461332.50	31.82

**STEP 2:** enter the how much data will be ingested each year

**STEP 3:** enter the percentage of data in the archive that will be exported each year

# Portability, Open Standards, Data Sovereignty





# Open Standards, Open Specifications, Open Source: Deployment



# Open Standards, Open Specifications, Open Source: Interaction



**RCLONE**



**XRootD**

**INVENIO**



**research  
object.org**



**Bagit & BDBags**

# Open Standards, Open Specifications, Open Source: Content



CEF Digital  
Connecting Europe



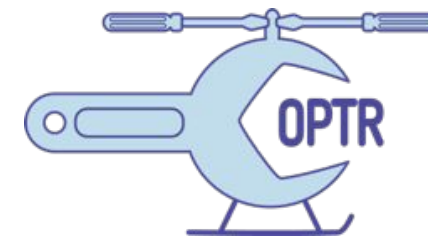
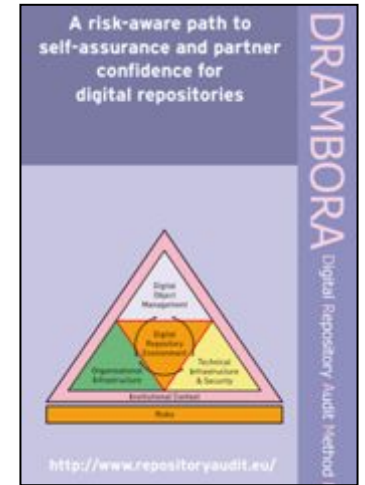


# **FAIR Forever, Digital Preservation, Long-term Access and Reuse**

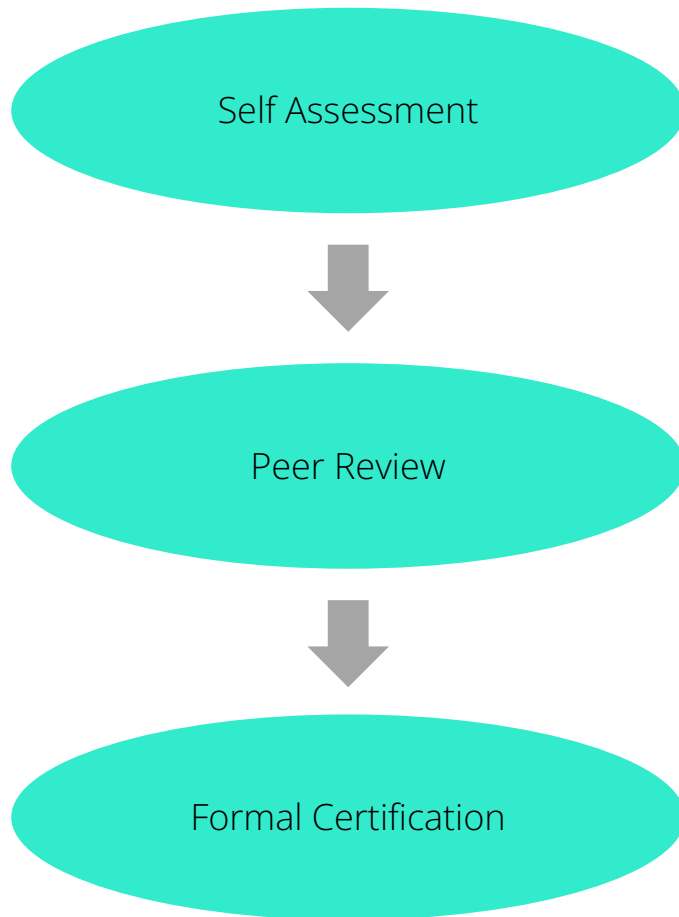




Functional Area	Level 1 (Show your contents)	Level 2 (Protect your contents)	Level 3 (Monitor your contents)	Level 4 (Sustain your contents)
Storage	Have two complete copies in separate locations. Document all storage media where content is stored. Put content into stable storage.	Have three complete copies with at least one copy in a separate geographic location. Document storage and storage media including the resources and dependencies they require to function.	Have at least one copy in a different storage media type. Track the obsolescence of storage and media.	Have at least three copies in geographic locations, each with a different disaster threat. Maximize storage diversification to avoid single points of failure.
Integrity	Verify integrity information if it has been provided with the content. Generate integrity information if not provided with the content. Track checks of content, include content for quarantine as needed.	Verify integrity information when moving or copying content. Use write blockers when working with original media. Back up integrity information and store copy in a separate location from the content.	Verify integrity information of content at least weekly. Document integrity information verification processes and outcomes. Perform audit of integrity information on demand.	Verify integrity information in response to specific events or activities. Repeat or repeat corrupted content as necessary.
Control	Describe the format and software agents that should be authorized to read, write, move, and delete content and apply those.	Document the format and software agents authorized to read, write, move, and delete content and apply those.	Monitor how and where the format and software agents that performed actions on content.	Perform periodic review of access/permissions logs.
Metadata	Create inventory of content, also documenting current storage locations. Backup inventory and store at least one copy separately from content.	Have enough metadata to know what the content is, who rights include usage restrictions of administrative, technical, descriptive, preservation, and structural.	Inventory what metadata standards to apply. Find and fill gaps in your metadata to meet those standards.	Record preservation actions associated with content and when those actions occur. Implement metadata standards often.
Content	Document the source and other essential content characteristics including how and when they were created.	Verify the format and other essential content characteristics. Build relationships with content creators to encourage sustainable practices.	Monitor for obsolescence and migration or obsolescence as when content is dependent.	Perform migration, normalization, emulation, and similar activities that ensure content can be accessed.



# Standards, External Assessment and Certification



**NDSA**



## FAIR Forever?

**Long Term Data Preservation Roles and Responsibilities, Final Report**

**February 2021 (V.7)**

Dr Amy Currie and Dr William Kilbride



EOSC FAIR Forever has received funding from the European Union under the EOSC Secretariat project. EOSCsecretariat.eu has received funding from the European Union's Horizon Programme call H2020-INFRAEOSC-05-2018-2019, grant Agreement number 831644. Europeana is an initiative of the European Union, financed by the European Union's Connecting Europe Facility and European Union Member States (<https://pro.europeana.eu/> and <https://www.europeana.eu/>)



# Environmental Sustainability



# Making LTDP more environmentally sustainable

- Keep less
  - e.g. appraisal, don't digitize everything
- Do less
  - e.g. minimum effort ingest, don't normalise
- Make smarter use of storage
  - e.g. deep archive, small footprint access copies
- Make more efficient use of IT resources
  - e.g. don't leave idle servers running
- Use environmentally friendly infrastructures
  - e.g. cloud with renewable energy

<https://www.dpconline.org/blog/is-digital-preservation-bad-for-the-environment>


## Toward Environmentally Sustainable Digital Preservation

Keith L. Pendergrass, Walker Sampson,  
Tim Walsh, and Laura Alagna

### ABSTRACT

Digital preservation relies on technological infrastructure (information and communication technology, ICT) that has considerable negative environmental impacts, which in turn threaten the very organizations tasked with preserving digital content. While altering technology use can reduce the impact of digital preservation practices, this alone is not a strategy for sustainable practice. Moving toward environmentally sustainable digital preservation requires critically examining the motivations and assumptions that shape current practice. Building on Goldman's challenge to current practices for digital authenticity and using Ehrenfeld's sustainability framework, we propose explicitly integrating environmental sustainability into digital preservation practice by shifting cultural heritage professionals' paradigm of appraisal, permanence, and availability of digital content.

The article is organized in four parts. First, we review the literature for differing uses of the term "sustainability" in the cultural heritage field: financial, staffing, and environmental. Second, we examine the negative environmental effects of ICT throughout the full life cycle of its components to fill a gap in the cultural heritage literature, which primarily focuses on the electricity use of ICT. Next, we offer suggestions for reducing digital preservation's negative environmental impacts through altered technology use as a stopgap measure. Finally, we call for a paradigm shift in digital preservation practice in the areas of appraisal, permanence, and availability. For each area, we propose a model for sustainable practice, providing a framework for sustainable choices moving forward.

© Keith L. Pendergrass, Walker Sampson, Tim Walsh, and Laura Alagna. 

### KEY WORDS

Digital preservation, Sustainability,  
Climate change, Appraisal, Permanence, Access

<https://dash.harvard.edu/handle/1/40741399>

# Cloud is greener than you might think

## Google Cloud Region Picker

This tool helps you pick a Google Cloud region considering carbon footprint, price and latency.

### Optimize for

Lower carbon footprint ②

Not important  Important

Lower price ②

Not important  Important

Lower latency ②

Not important  Important

### Where is your traffic coming from?

Ukraine

United Arab Emirates

United Kingdom

United States

Uruguay

### Recommended regions

europe-north1  
Hamina, Finland

- Carbon Free Energy: 94%
- Grid carbon intensity: 133 gCO<sub>2</sub>eq/kWh
- 1. Google Compute Engine price: \$0.024016 / vCPU-hour

us-central1  
Iowa, USA

- Carbon Free Energy: 93%
- Grid carbon intensity: 454 gCO<sub>2</sub>eq/kWh
- 2. Google Compute Engine price: \$0.021811 / vCPU-hour

northamerica-northeast1  
Montréal, Canada

- Grid carbon intensity: 27 gCO<sub>2</sub>eq/kWh
- 3. Google Compute Engine price: \$0.024013 / vCPU-hour



<https://www.google.com/about/datacenters/gallery/#hamina-exterior-landscape>



<https://www.google.com/about/datacenters/gallery/#st-ghislain-solar-panels>

<https://cloud.withgoogle.com/region-picker/>



# Cloud providers can achieve very high energy efficiency



**ENERGY**

## Recalibrating global data center energy-use estimates

Growth in energy use has slowed owing to efficiency gains that smart policies can help maintain in the near term

By **Eric Masanet**<sup>1,2</sup>, **Arman Shehabi**<sup>3</sup>,  
**Nuoa Lei**<sup>2</sup>, **Sarah Smith**<sup>3</sup>, **Jonathan Koomey**<sup>4</sup>

**D**ata centers represent the information backbone of an increasingly digitalized world. Demand for their services has been rising rapidly (1), and data-intensive technologies such as artificial intelligence, smart and connected energy systems, distributed manufacturing systems, and autonomous vehicles promise to increase demand further (2). Given that data centers are energy-intensive enterprises, estimated to account for around 1% of worldwide electricity use, these trends have clear implications for global energy demand and must be analyzed rigorously. Several oft-cited yet simplistic analyses claim that the energy used by the world's data centers has doubled over the past decade and that their energy demand for data center services rises rapidly, so too must their global energy use. But such extrapolations based on recent service demand growth indicators overlook strong countervailing energy efficiency trends that have occurred in parallel (see the first figure). Here, we integrate new data from different sources that have emerged recently and suggest more modest growth in global data center energy use (see the second figure). This provides policy-makers and energy analysts a recalibrated understanding of global data center energy use, its drivers, and near-term efficiency potential.

Assessing implications of growing demand for data centers requires robust understanding of the scale and drivers of global data center energy use that has eluded many policy-makers and energy analysts. The reason for this blind spot is a historical lack of "bottom-up" information

As demand for data centers rises, energy efficiency improvements to the IT devices and cooling systems they house can keep energy use in check.

Bottom-up analyses tend to best reflect this broad range of factors, generating the most credible historical and near-term energy-use estimates (7). Despite several recent national studies (8), the latest fully replicable bottom-up estimates of global data center energy use appeared nearly a decade ago. These estimates suggested that the worldwide energy use of data centers had grown from 153 terawatt-hours (TWh) in 2005 to between 203 and 273 TWh by 2010, totaling 1.1 to 1.5% of global electricity use (9).

Since 2010, however, the data center landscape has changed dramatically (see the first figure). By 2018, global data center workloads and compute instances had increased more than sixfold, whereas data center internet protocol (IP) traffic had increased by more than 10-fold (1). Data center storage capacity has also grown rapidly, increasing by an estimated factor of 25 over the same time period (1, 8). There has been a tendency among analysts to use such service demand trends to simply extrapolate earlier bottom-up energy values, leading to unreliable predictions of current and future global data center energy use (3–5). They might, for example, scale up previous bottom-up values (e.g., total data center energy use in 2010) on the basis of the growth rate of a service demand indicator (e.g., growth in global IP traffic from 2010 to 2020) to arrive at an estimate of future energy use (e.g., total data center energy use in 2020).

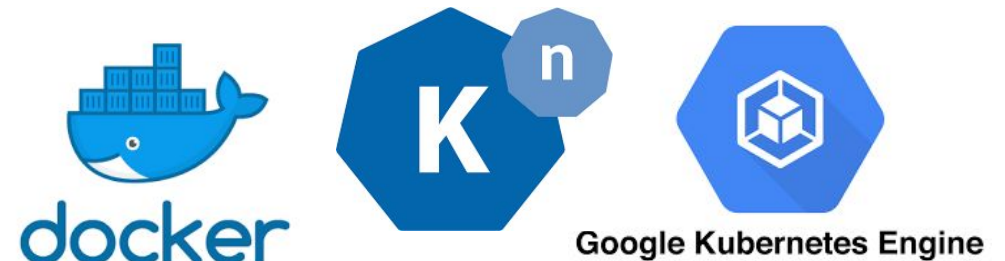
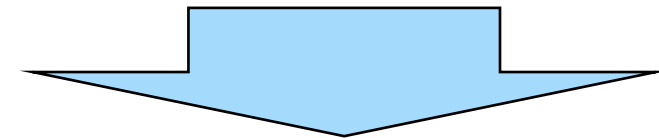
But since 2010, electricity use per computation of a typical volume server—the workhorse of the data center—has dropped by a factor of four, largely owing to processor efficiency improvements and reductions in idle power (10). At the same time, the watts per terabyte of installed storage has dropped by an estimated factor of nine owing to storage-drive density and efficiency gains (8). Furthermore, growth in the number of servers has slowed considerably owing to a fivefold increase in the average number of compute instances hosted per server (owing to virtualization), alongside steady reductions in data center power usage effectiveness (PUE, the total amount

Downloaded from <http://science.sciencemag.org/> on March 20, 2020

[https://datacenters.lbl.gov/sites/default/files/Masanet\\_et\\_al\\_Science\\_2020.full\\_.pdf](https://datacenters.lbl.gov/sites/default/files/Masanet_et_al_Science_2020.full_.pdf)



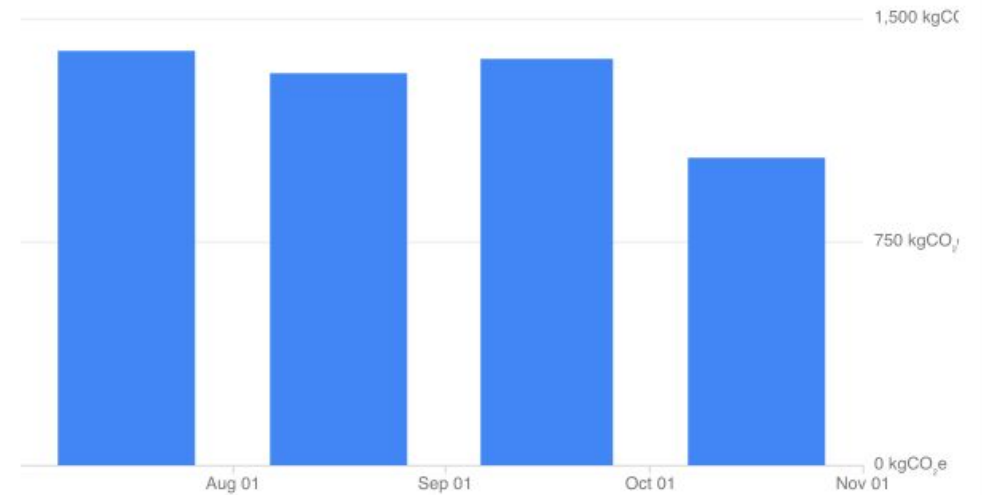
<https://www.stickermule.com/marketplace/3442-there-is-no-cloud>



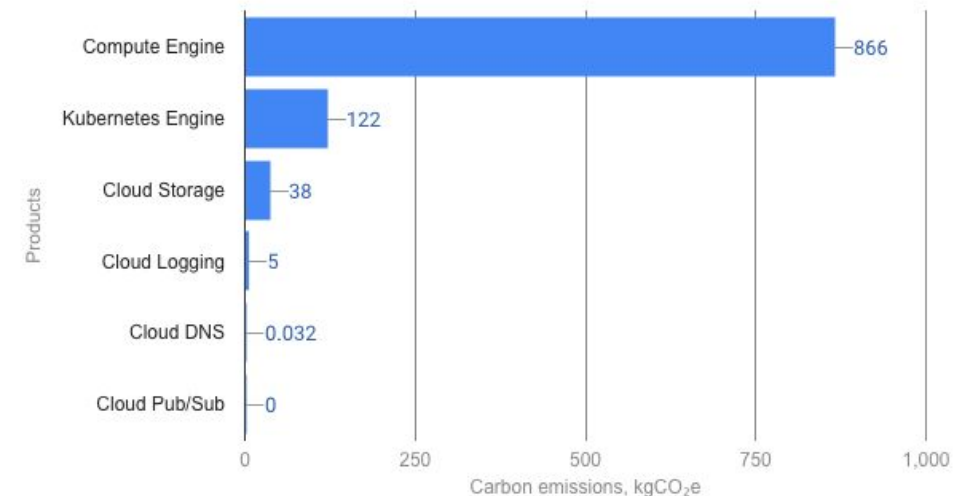
# Strategy for energy efficiency and low carbon footprint

- Users can choose what processing to do on their content
- Deep archive storage for infrequently accessed data
- Serverless computing: only consume resources when needed
- Deployment into energy efficient data centres
- Allow applications to run near to the data

Gross monthly carbon emissions



Gross carbon emissions by product in October 2021



# Summary





# Summary

- The scale of scientific datasets mean that new approaches to LTDP are necessary
- Automation, microservices, serverless computing and cloud IaaS are powerful combination
- Efficiency and costs can be measured, managed and optimized (€ and carbon)
- Reduced consumption and choice over cloud location/provider helps environmental sustainability
- Open standards, open specifications and open source are key to portability and interoperability





arkivum

Bringing archived data to life



QUESTIONS?



[matthew.addis@arkivum.com](mailto:matthew.addis@arkivum.com)



[orcid.org/0000-0002-3837-2526](https://orcid.org/0000-0002-3837-2526)



[www.arkivum.com](http://www.arkivum.com)

[www.arkivum.com](http://www.arkivum.com)

[hello@arkivum.com](mailto:hello@arkivum.com)



ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

Libnova - CSIC - University of Barcelona -  
Giaretta Associates - AWS - Voxility - Bidaidea



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.





# PILOT PHASE KICK-OFF EVENT

2021-11-29





**Big thanks to the Archiver team, EU Representatives and the LIBNOVA consortium team.**



# ARCHIVER

ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

## INTERVIEW WITH THE SELECTED BIDDERS

libnova

 **CSIC**  
CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

 UNIVERSITAT DE  
BARCELONA

 **Giaretta  
Associates**

**aws**

**voxility**

 **bidaidea**

# Public Award Ceremony





# LABDRIVE Core capacities



# How are you and the ARCHIVER project benefiting LIBNOVA?



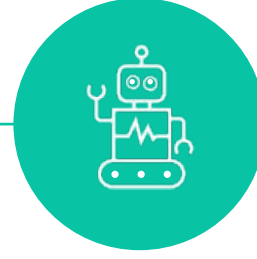
## Market

### Time to market:

Shorter development lifecycle leading to a faster time to market (5 to 2 years)

### Market visibility:

Several EU/USA Universities and a large European pharmaceutical signing contracts.



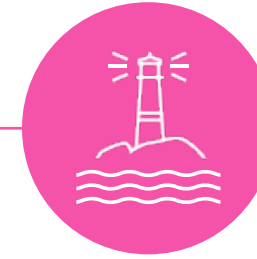
## Product

### Accelerated innovation:

Being able to work with the 4 buyers maximizes success chances in first iteration.

### Standards:

We are bringing several – previously unknown – standards onboard.



## Team

### EU<sup>nt</sup>husiasm:

LIBNOVA consortium's team feels enthusiastic about being part of something larger, relevant to the EU. The EU is leading the R&D in this field.

### Happiness:

Everyone have been working really hard, but really happy about it!

# How are we benefiting the community?

## Making the practice more efficient

- Direct: Optimized storage costs, low operational costs.
- Indirect: Less time/resources for producers/researchers to use it.

## Making it available to a broader audience

- Demystifies preservation. Easy to understand and to use.
- Opens the practice to large volume datasets and to more advanced organizations.

## Making it easier to apply best

- Fully conformant to ISO 14721 and ISO16363 (and several other standards)
- Full support for FAIR/TRUST data models, workflows, etc.

practices



# How are we benefiting the community?

## Reducing the environmental impact of the community preservation activities

- Reuse what you have: on-premise deployments using existing infrastructure are possible.
- Consume only when needed: Resources are consumed when there is something to do. Scaling from 36 Kubernetes pods to ~5000 in 32 minutes. Process the workload and then back to 36 pods.
- “Environment impact” topic is included in every architecture specification/analysis. The team is proud of many small and smart optimizations. For example, the hashing algorithm.
- Using carbon-neutral providers and data centres.

## Contributing to the European digital sovereignty and digital development and leadership

- Cloud provider independence
- Decouples content from providers
- Open to the emergence of EU-based cloud providers.

# Thanks again!



Any questions?





ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

# Closing remarks

João Fernandes (CERN)



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.

# Closing Remarks

## Upcoming Milestones for Buyers / Early Adopter organisations

- Today: Pilot kick-off meeting
- January 2022: Full capacity pilot platforms available
- April - June 2022: Project Review meetings with demos of the resulting services
  - Webinars for training administrators and end users in the Pilot platforms

## Upcoming Public Events (February - June 2022)

- Industry related webinars about the advantages of PCP projects - February 2022
  - Procurement of research and development of new innovative solutions before they are commercially available
- Thematic webinars for potential Early Adopters of ARCHIVER - March-May 2022

**Dates and Registration to be announced at <https://archiver-project.eu/>**



ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

# Thank you!

*Join our community!*



[www.archiver-project.eu](http://www.archiver-project.eu)



[@ArchiverProject](https://twitter.com/ArchiverProject)



[company/archiver-project](https://company/archiver-project)



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.