ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

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*

***Starting Date**: 1st of January 2019*

***Duration**: 42 Months*

***Coordinator**: CERN (Lead Procurer)*

European Commission

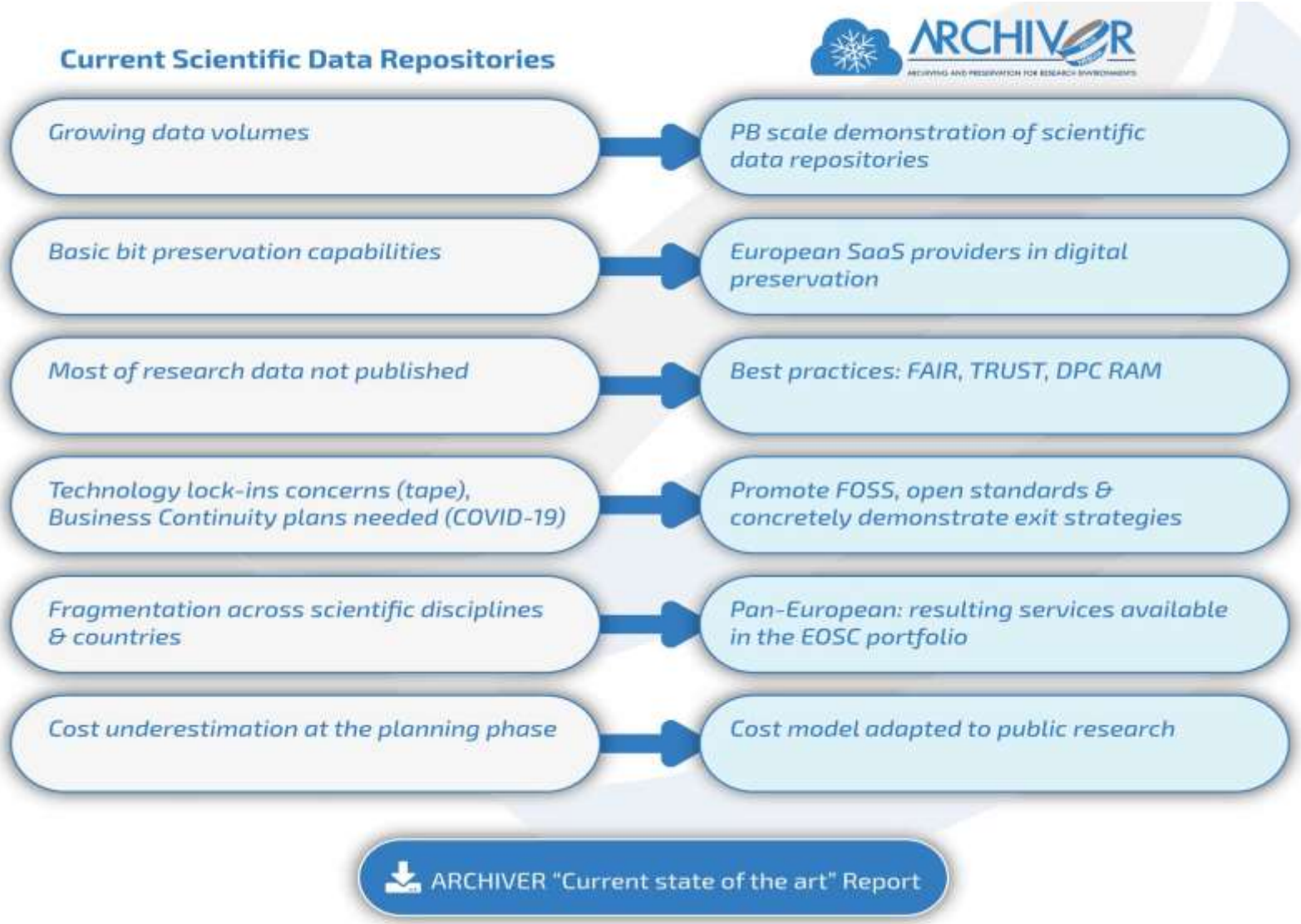
Buyers Group (BG) - Public organisations committing funds to contribute to a joint-R&D-procurement, research data use cases and R&D testing effort.

PHOTON-NEUTRON SCIENCES · ASTROPHYSICS & COSMOLOGY · LIFE SCIENCES · HIGH ENERGY PHYSICS · LONG TAIL OF SCIENCE

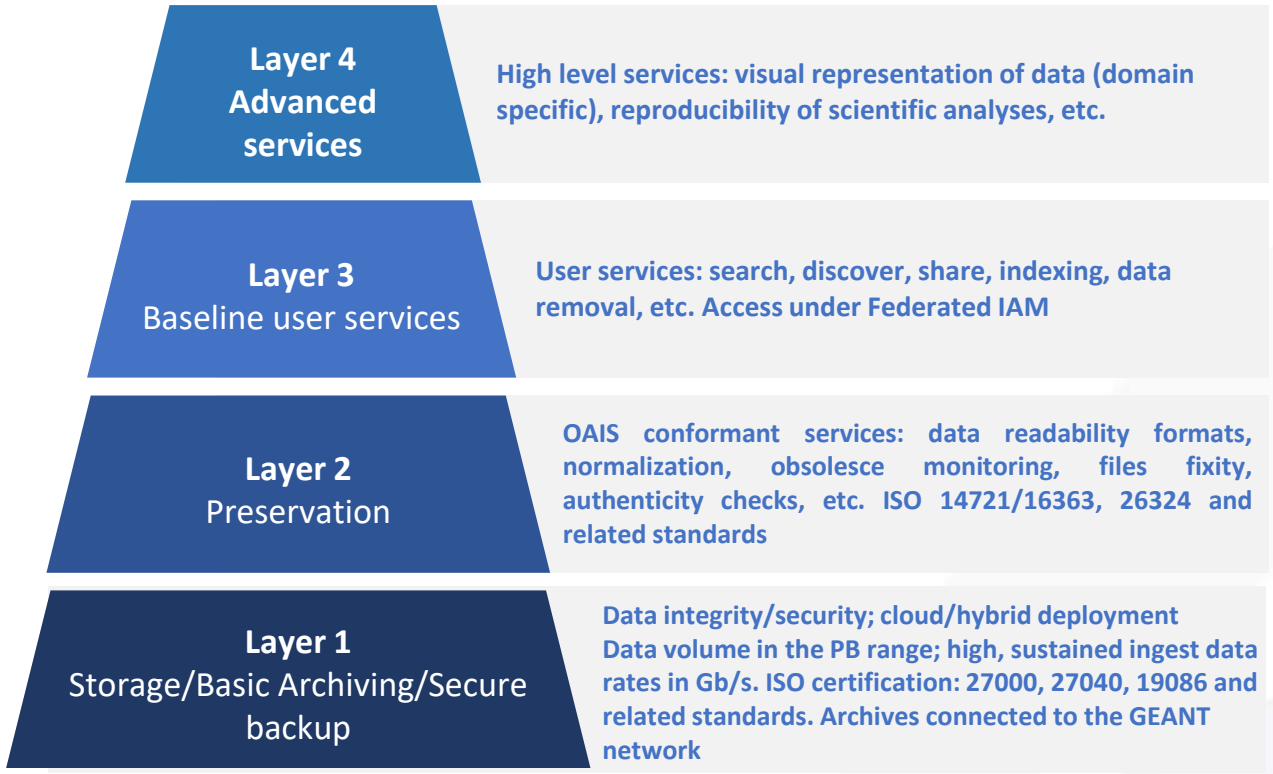
Buyers + Experts → Consortium

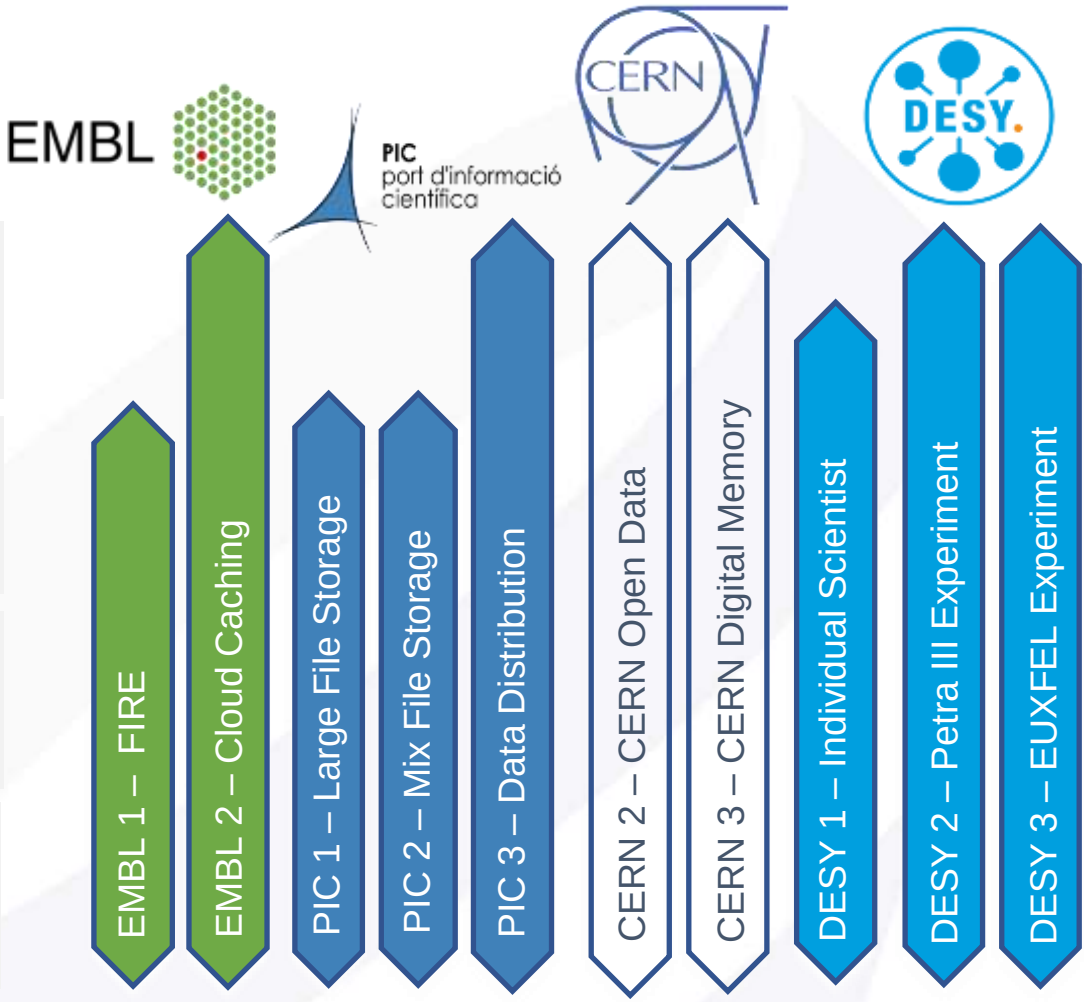
Experts - Partner organisations bringing expertise in requirement assessment and promotion activities

Problem and Solution



R&D - Scope

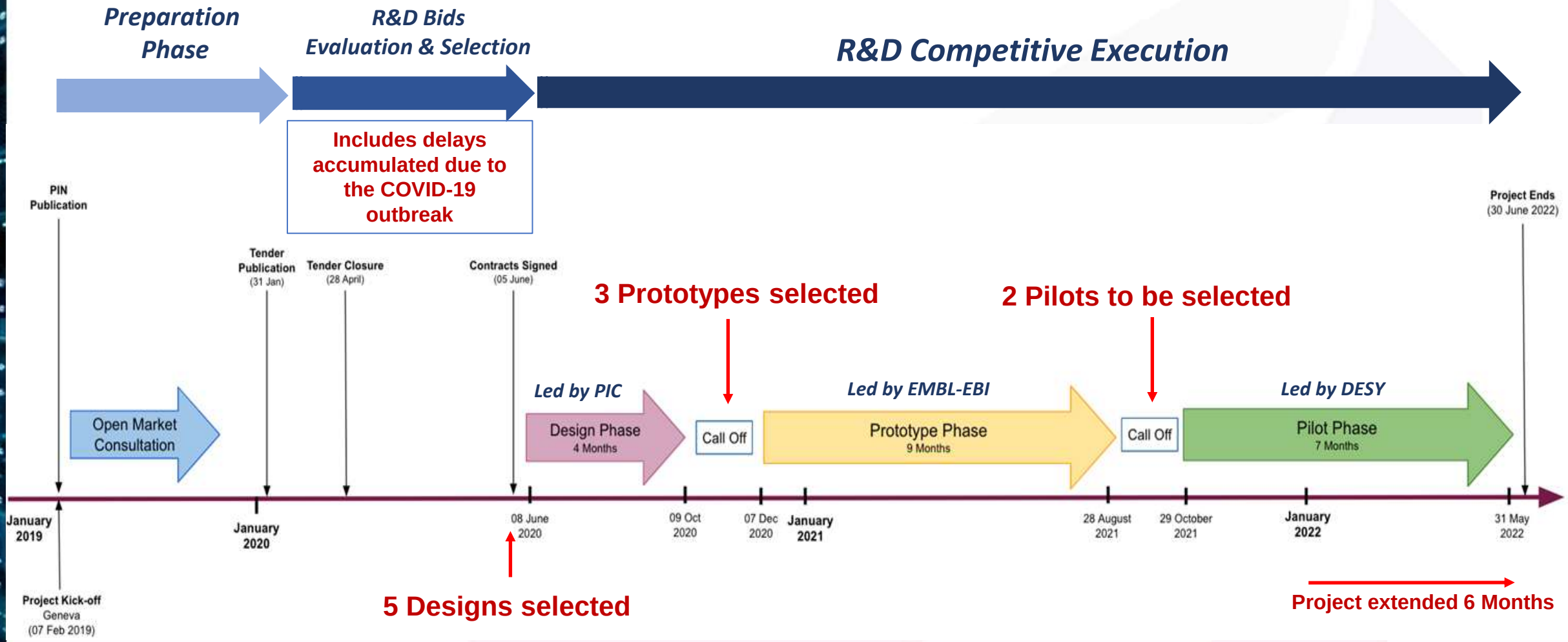




Scientific use cases deployments documented at: https://www.archiver-project.eu/deployment-scenarios

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

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

On behalf of:

ARCHIVER Consortium (CERN, EMBL-EBI, DESY, PIC, Addestino & TRUST-IT)

& **ARCHIVER Pilot Phase Lead Contractors** (Arkivum, Libnova)

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.

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

Publication in Zenodo mid-December

ARCHIVER

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?

ARCHIVER

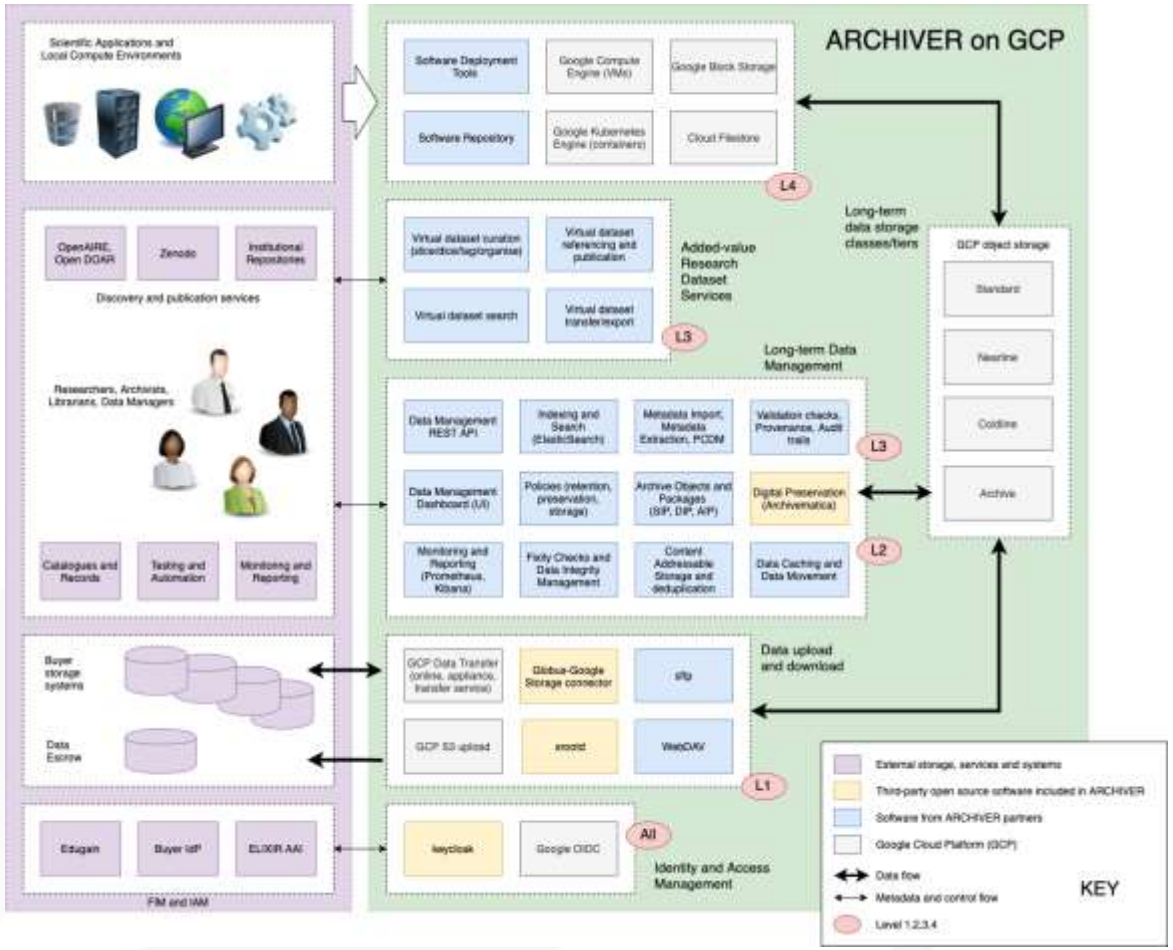
ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

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**

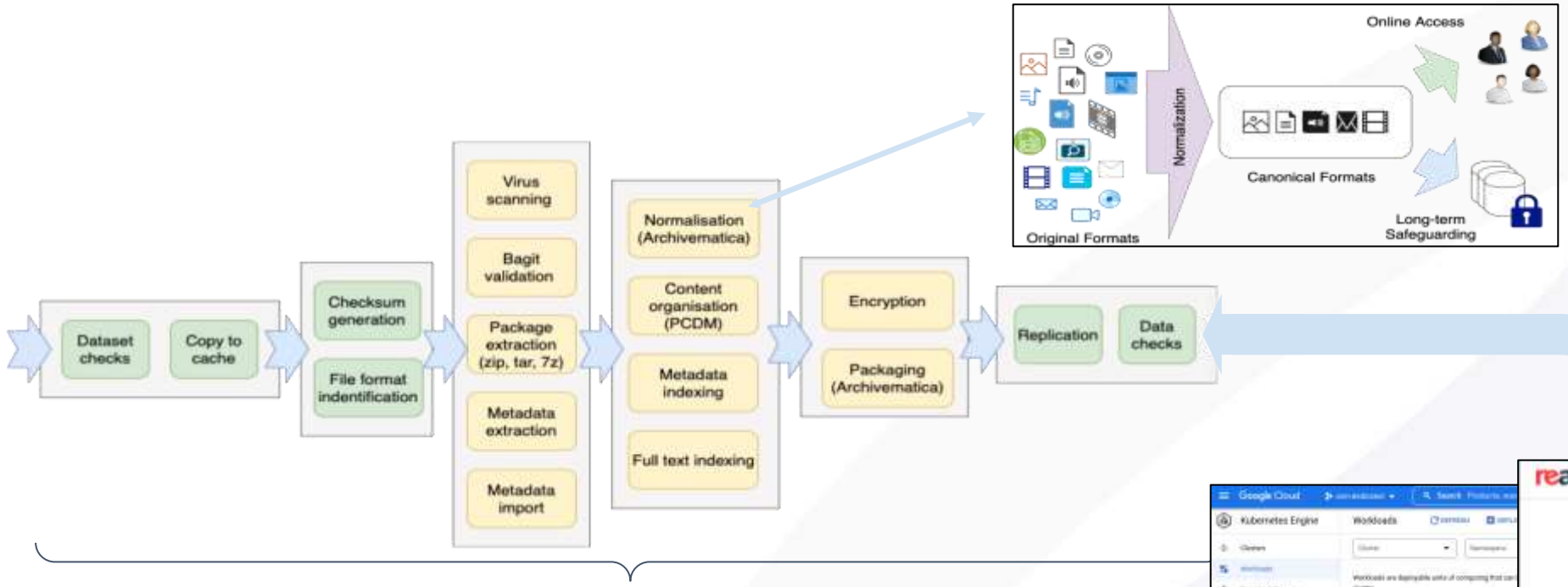
- SaaS solution in the cloud or on-premise, provided as a managed service
- Autoscaling based on Kubernetes, supporting multi-petabyte volumes of billions of objects
- Certified ISO 9001 / ISO 27001
- Solution supports NDSA Levels of Preservation, DPC RAM, CoreTrustSeal and FAIR principles, with mappings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



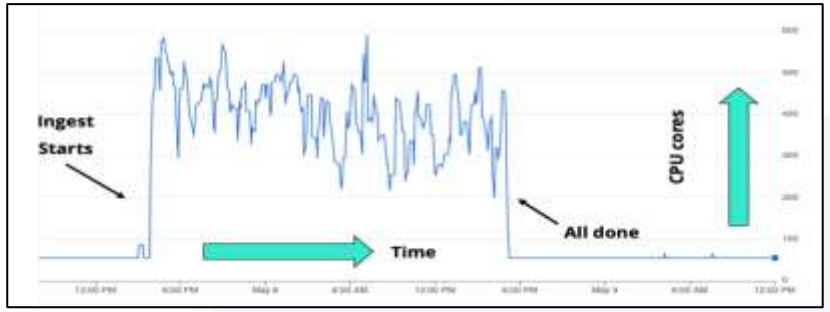
Prototype architecture of the Arkivum consortium (image courtesy of the Arkivum consortium)

Levels of Digital Preservation

Selected Consortia: **Arkivum**



Example of ingest and archiving of Astronomy Dataset:
440k image files ingested as 2789 big data bags **within 24 hours**

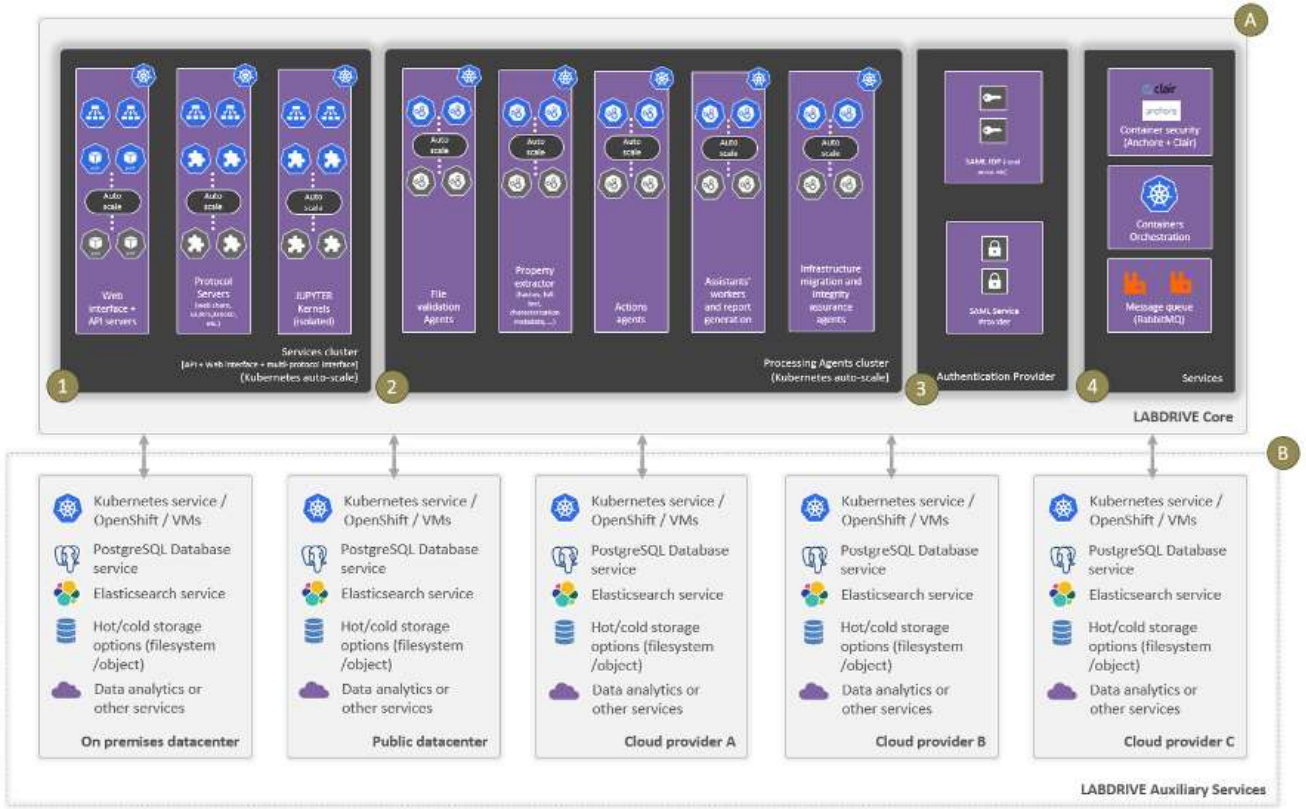


Example of analysis in GCP of CERN Open Data datasets

Example of dataset landing pages created in InvenioRDM for research datasets in the Arkivum LTDP solution

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



Prototype architecture of the LIBNOVA Consortium (image courtesy of the LIBNOVA Consortium)

Scalable platform:

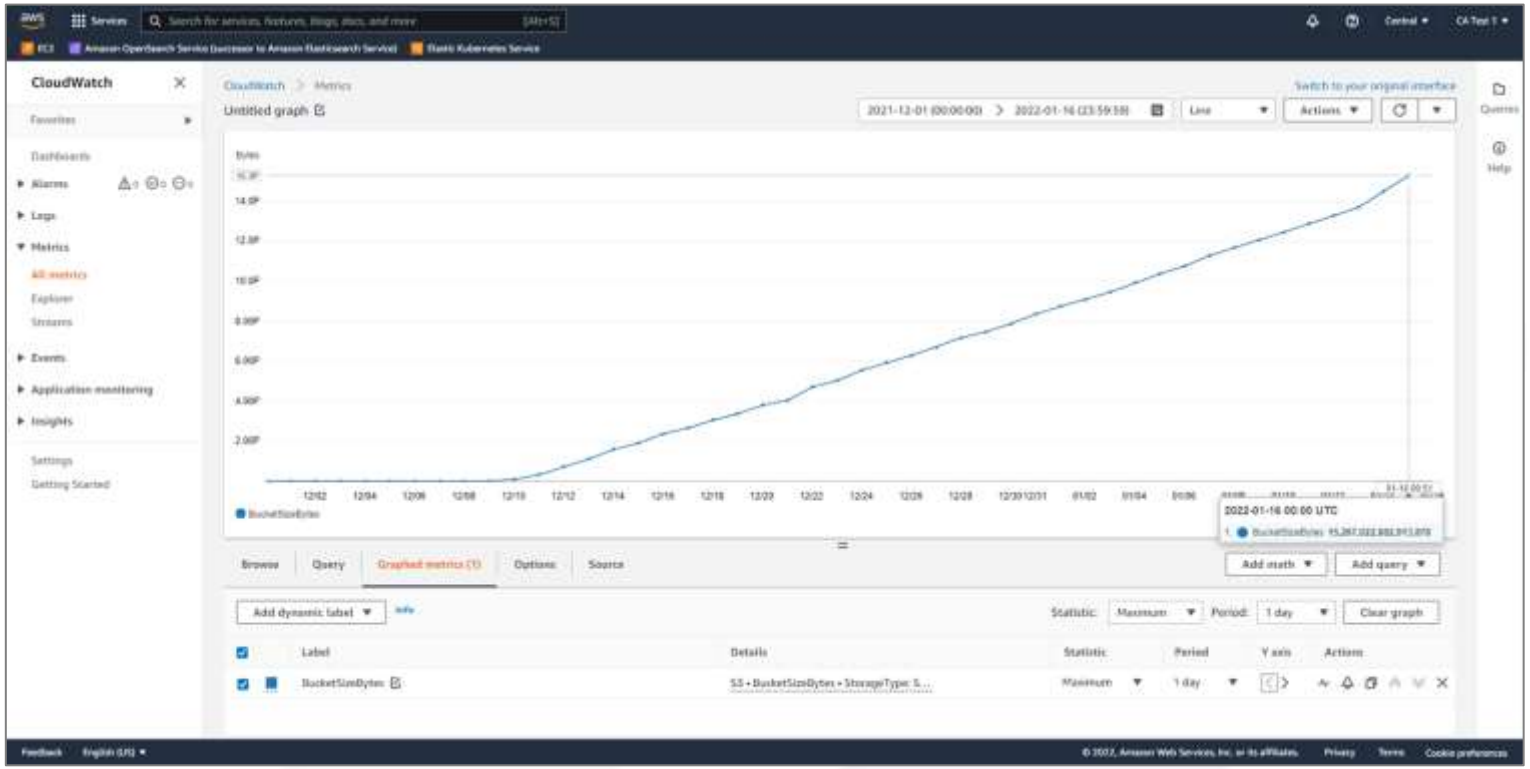
15

Petabytes

Ingested and preserved in **one month**

500TB

Ingested every day



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.

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.)