ARCHIVER

ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

Towards environmentally sustainable long-term digital preservation

IDCC 2022

Ignacio Peluaga (CERN)

June 14th 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*

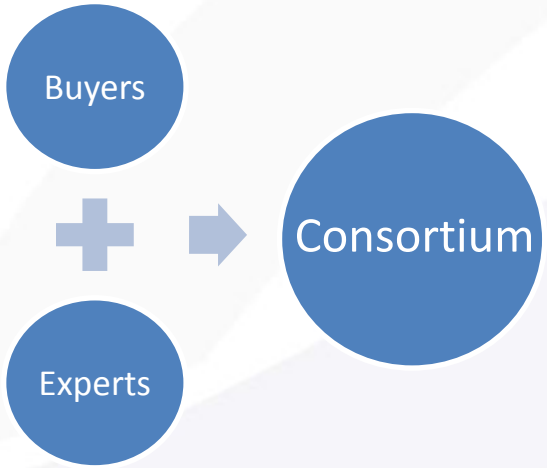
***Coordinator**: CERN (Lead Procurer)*

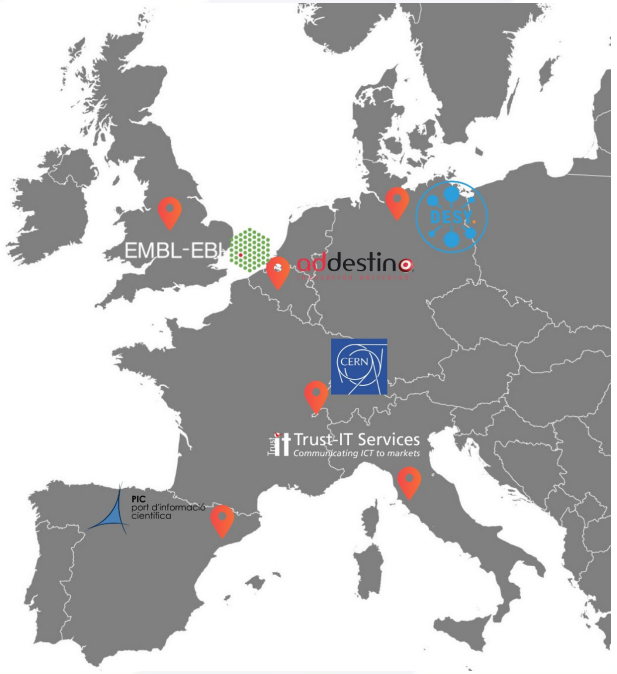
European Commission

EUROPEAN OPEN SCIENCE CLOUD

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

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



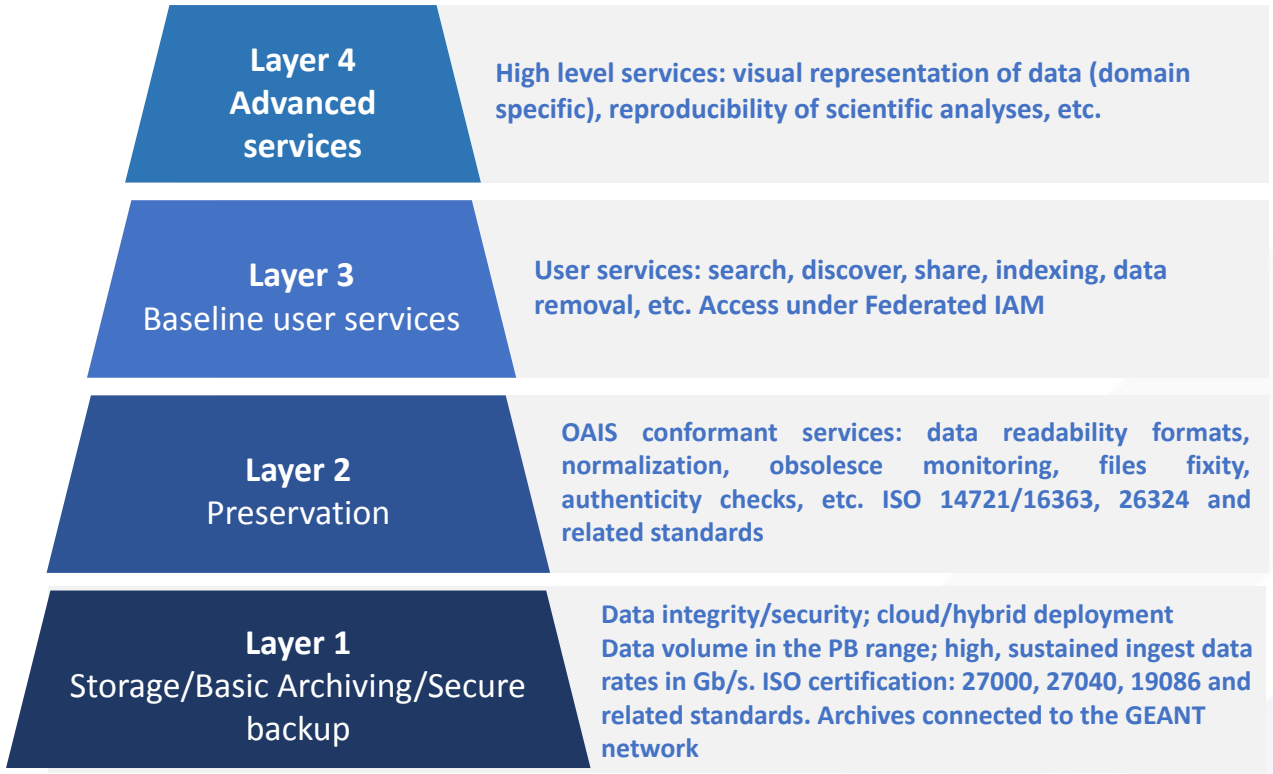


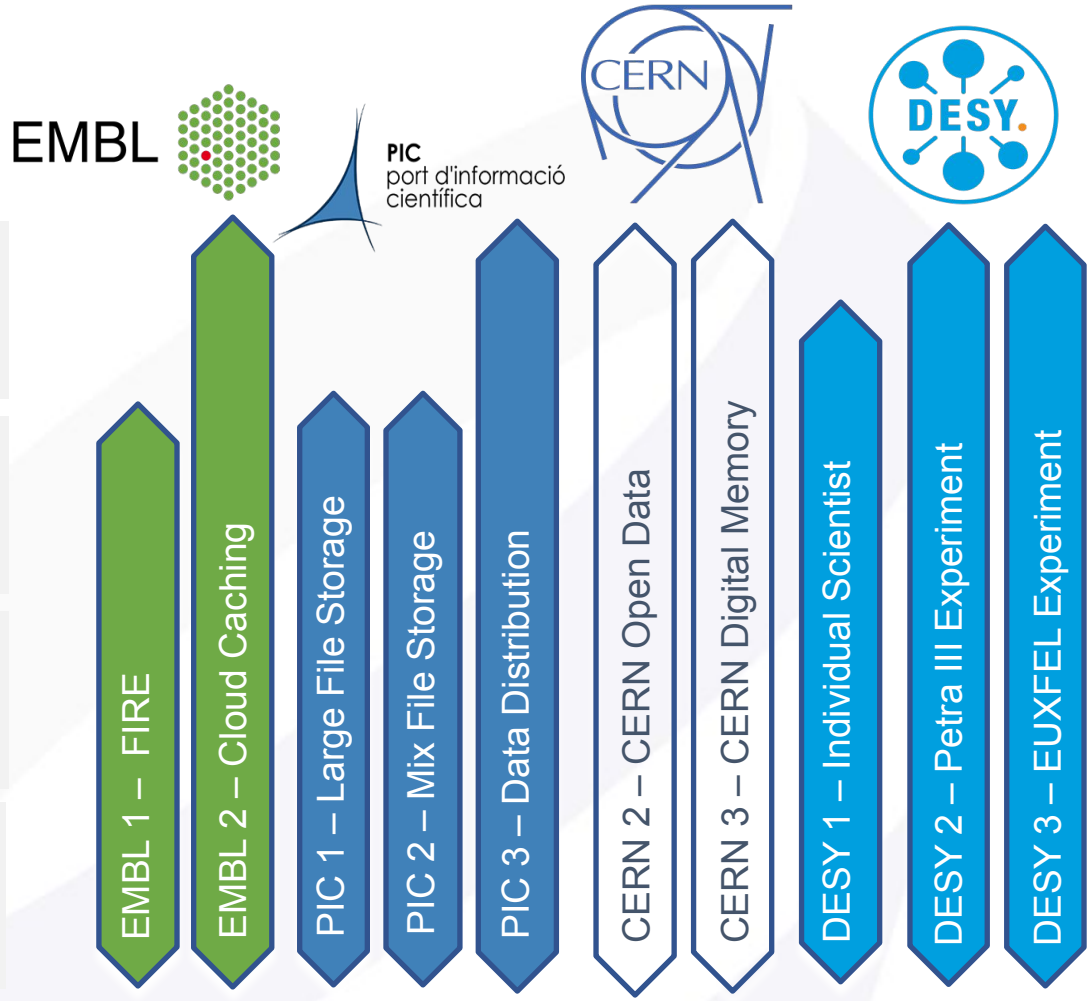
Progress Beyond the state of the art

Scientific Data Repositories before ARCHIVER		ARCHIVER
Growing data volumes	→	PB scale demonstration of scientific data repositories
Basic bit preservation capabilities	→	European SaaS providers in digital preservation
Most of research data not published	→	Best practices: FAIR, TRUST, DPC RAM
Technology lock-ins concerns (tape), Business Continuity plans needed (COVID-19)	→	Promote FOSS, open standards & demonstrate exit strategies
Fragmentation across scientific disciplines & countries	→	Pan-European: resulting services to be available in the EOSC portfolio
Cost underestimation at the planning phase	→	Cost-effective and environmentally sustainable model adapted to public research

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

R&D - Use Cases

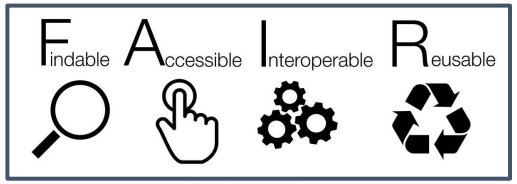




Scientific use cases deployments documented at: https://www.archiver-project.eu/deployment-scenarios

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

FAIR assessment



FAIRSFAIR
Fostering Fair Data Practices in Europe

F-UJI
Automated FAIR Data Assessment Tool

Process in three steps:

1. **Ingest datasets**

Upload/ingest datasets from the buyers with DataCite or DublinCore metadata

1. **Perform automated FAIR assessment**

Done with FAIRsFAIR developed F-UJI tool: https://github.com/pangaea-data-publisher/fuji

1. **Complement with manual assessment**

The contractors provided reports on how their platforms can keep data FAIR

R&D Execution

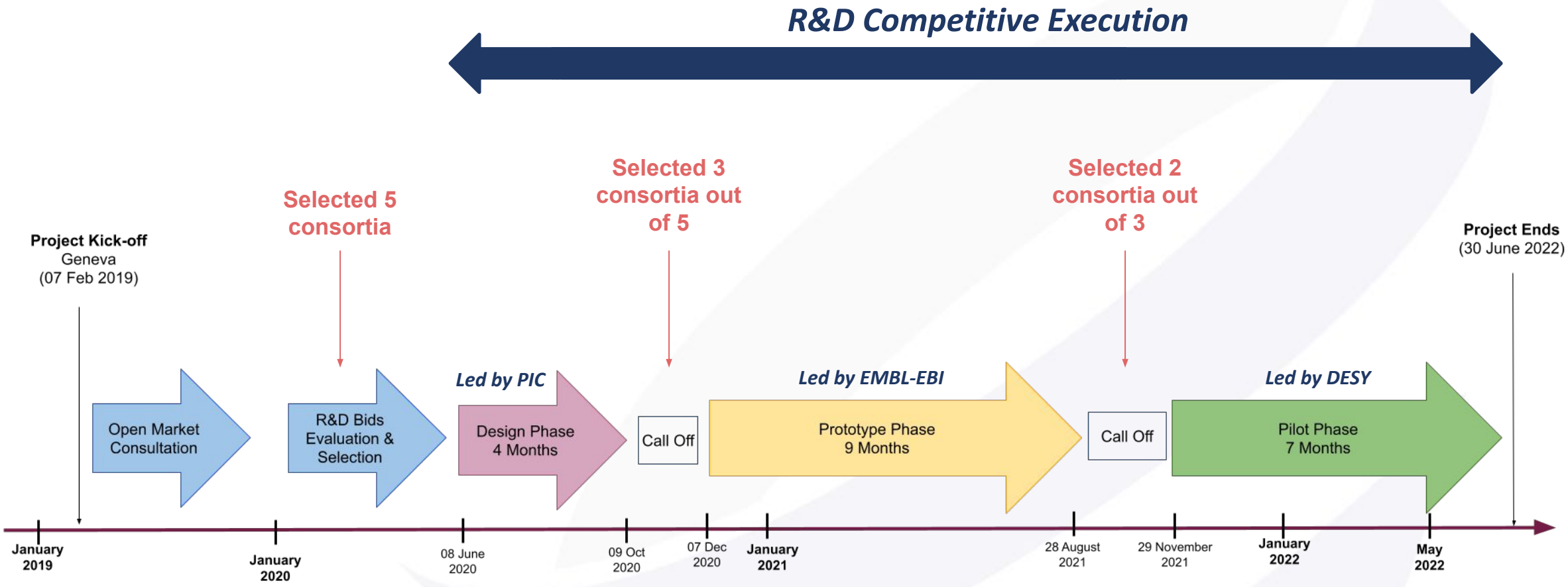
ARCHIVER is following an implementation on **three** phases with multiple competing **consortia**:

- **Phase 1 - Solution Design**
  - 5 selected consortia (https://archiver-project.eu/design-phase-award) develop designs including architecture and technical components
  - The activity during this phase has produced the results to be taken into account in the selection process that allows a consortium to proceed to the subsequent project phase

- **Phase 2 - Prototype Development**
  - 3 selected consortia from Phase 1 (https://archiver-project.eu/prototype-phase-award) build prototypes based on the designed solutions
  - Make them available to the buyers group for testing purposes

- **Phase 3 - Pilot Deployment**
  - 2 selected consortia from Phase 2 (https://archiver-project.eu/pilot-phase-award) deploy expanded prototype services
  - These services will potentially be exposed to end-users and early adopters, to determine if they are suitable for their needs

Project Timeline



Pilot Phase Selected Consortia

arkivum
Bringing archived data to life

Google Cloud

libnova

CSIC
Consejo Superior de Investigaciones Científicas

aws

Universitat de Barcelona

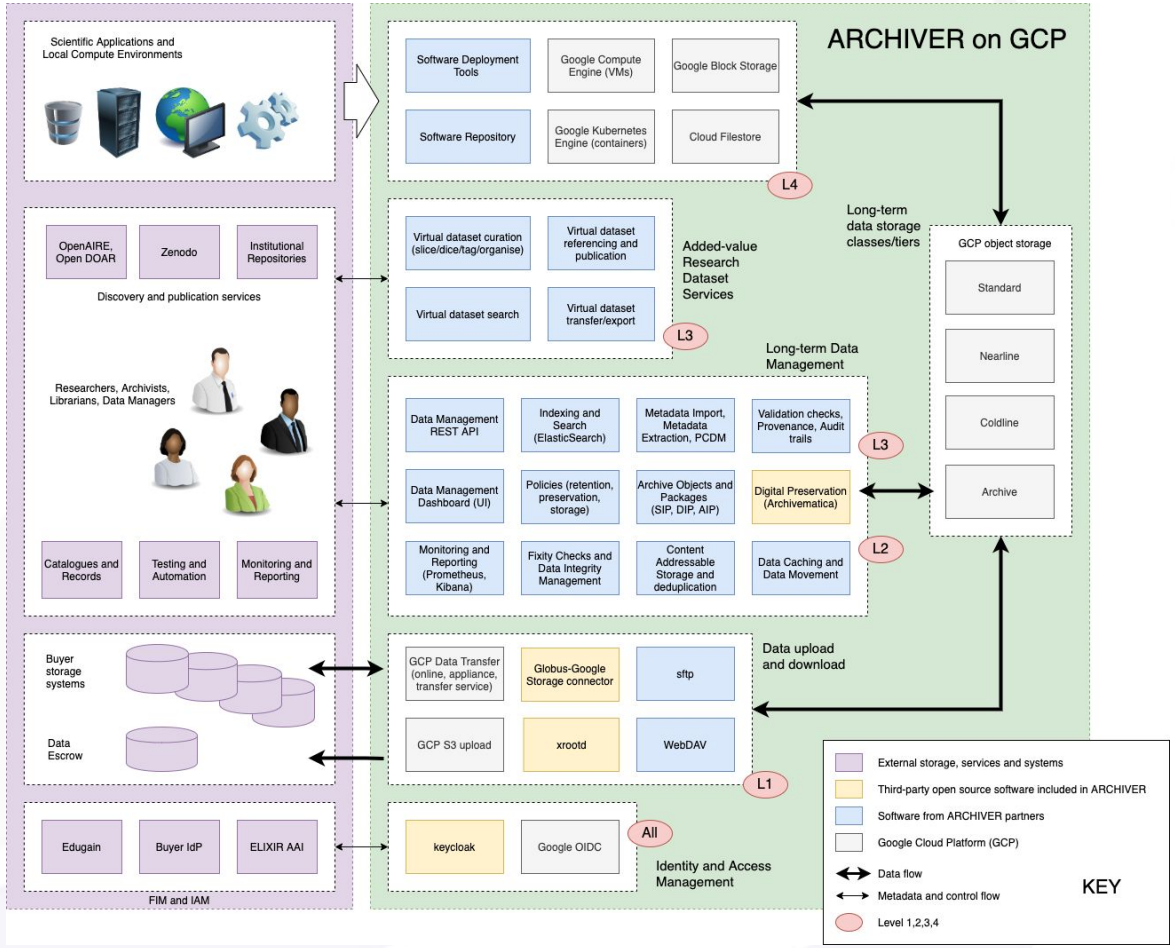
voxility

bidaidea

Giaretta Associates

Selected Consortia: **Arkivum**

- Overall architecture composed of micro-services to **scale** to multi-petabyte volumes of billions of objects
- Based on Kubernetes: autoscales, meaning no idle resources which **reduces** costs and carbon emissions
- Different storage options, for example deep archive/cold storage for infrequently accessed data = **cheaper**
- Google Cloud Platform (GCP) Infrastructure **carbon neutral** since 2007, with multiple low carbon data centers in Europe: carbon free by 2030*



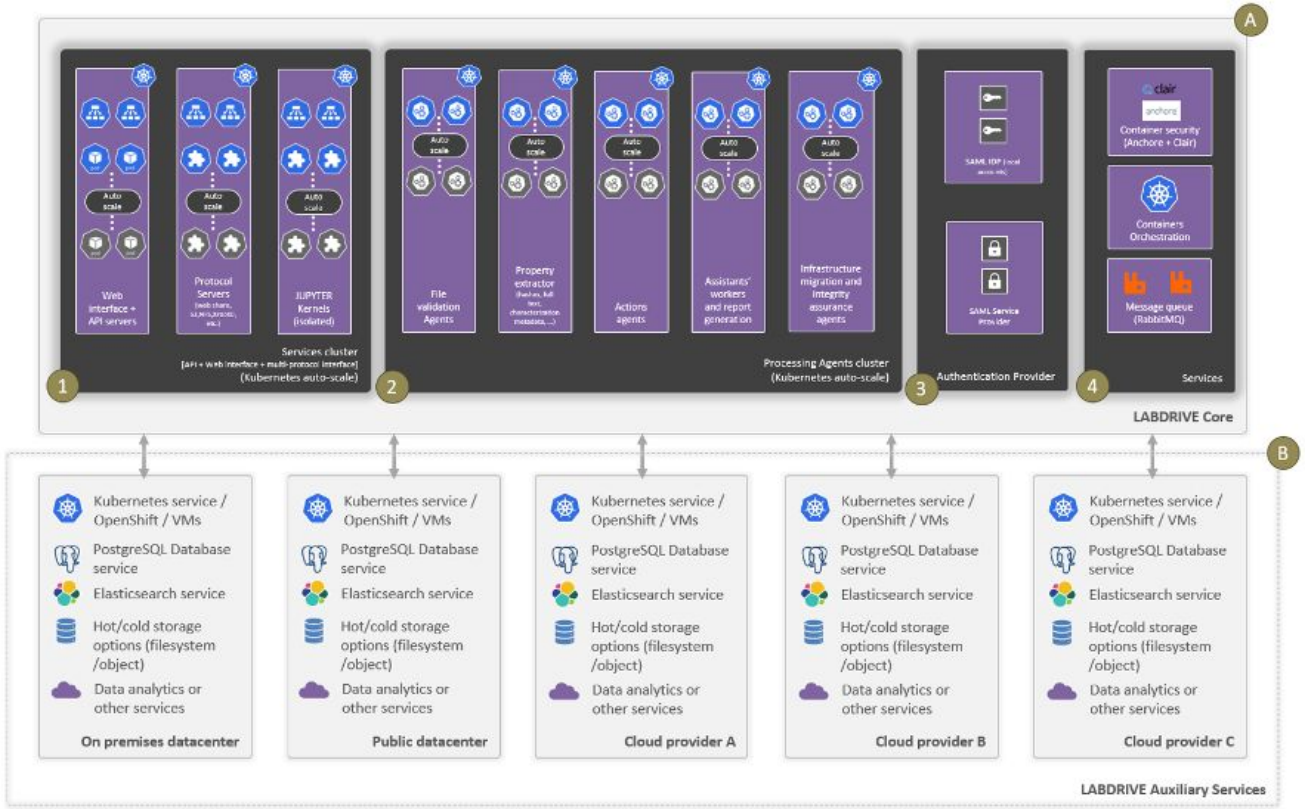
Prototype architecture of the Arkivum consortium (image courtesy of the Arkivum consortium)

*https://cloud.google.com/sustainability

Selected Consortia: **Libnova**

- Prototype based on LibSAFE SaaS
- Using infrastructure provided by AWS that aims to use **only renewable power** by 2025*, can be deployed **on-premises** too
- Runs on Kubernetes, fully scalable: adjustable number of components based on service demand which translates to cost and environmental **effectiveness**:
    - *Scaling from 36 Kubernetes pods to ~5000 in 32 minutes. Process the workload and then back to 36 pods.*
- QoS **optimization** of storage tiers considering carbon emissions among other factors. Less frequently accessed is cheaper

*https://aws.amazon.com/energy/sustainability/



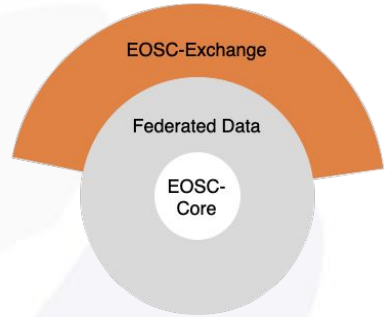
Prototype architecture of the Libnova consortium (image courtesy of the Libnova consortium)

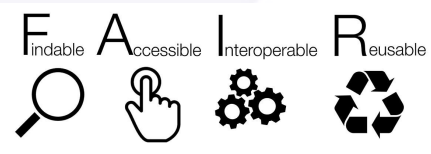
An example of autoscaling

Image courtesy of the Arkivum consortium

Conclusions

- 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
- ARCHIVER is acting as a template to **commoditise** archiving and preservation at scale in research domains
- ARCHIVER is promoting a **sustainable model** with services that will exist beyond the project lifetime in the context of the **EOSC**:
  - Cloud providers can achieve very high energy efficiency. On-prem possible as well “reuse what you have”
  - Automation, microservices, serverless computing and cloud IaaS are powerful combination: make more efficient use of resources e.g. don’t leave idle servers running
  - Choice over cloud location/provider helps environmental sustainability: use environmentally friendly infrastructures e.g. clouds with renewable energy
  - Open standards, open specifications and open source are key to portability and interoperability
  - Make smarter use of storage e.g. deep/cold/infrequent access archive, small footprint access copies
- ARCHIVER’s last phase -*Pilot Phase*- coming to an end but future plans and activities are already underway





ARCHIVER

ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

Thank you! Questions?

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-hWmQatqdIhIzw