



ARCHIVER

ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

Archiving & Preservation for Research Environments

Environmental Considerations

Digital Preservation Coalition

Environmentally sustainable digital preservation - moving from theory to practice

João Fernandes (CERN) [joao.fernandes AT cern.ch](mailto:joao.fernandes@cern.ch)

11th November 2021



Digital **Preservation** Coalition



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)



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



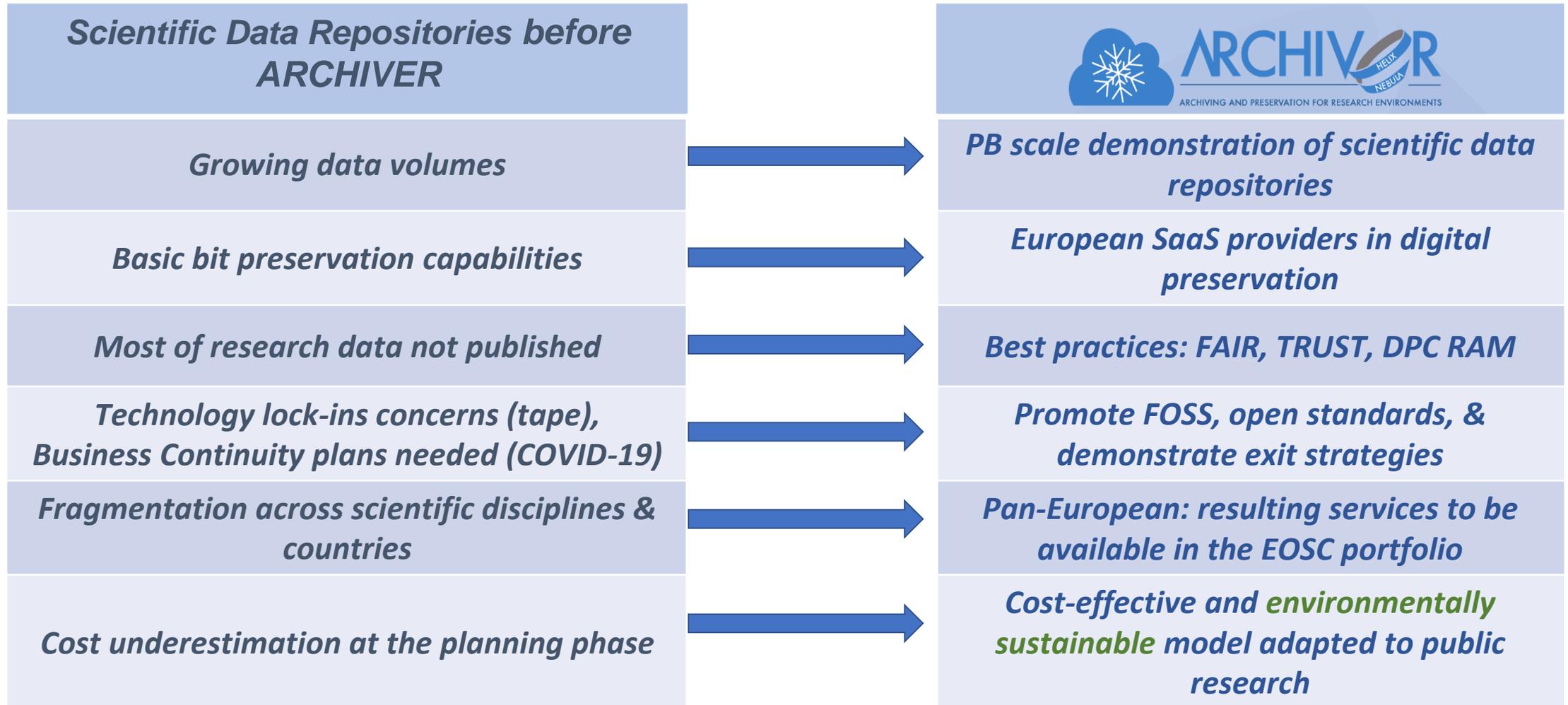
Experts - Partner organisations bringing expertise in requirement assessment and promotion activities



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

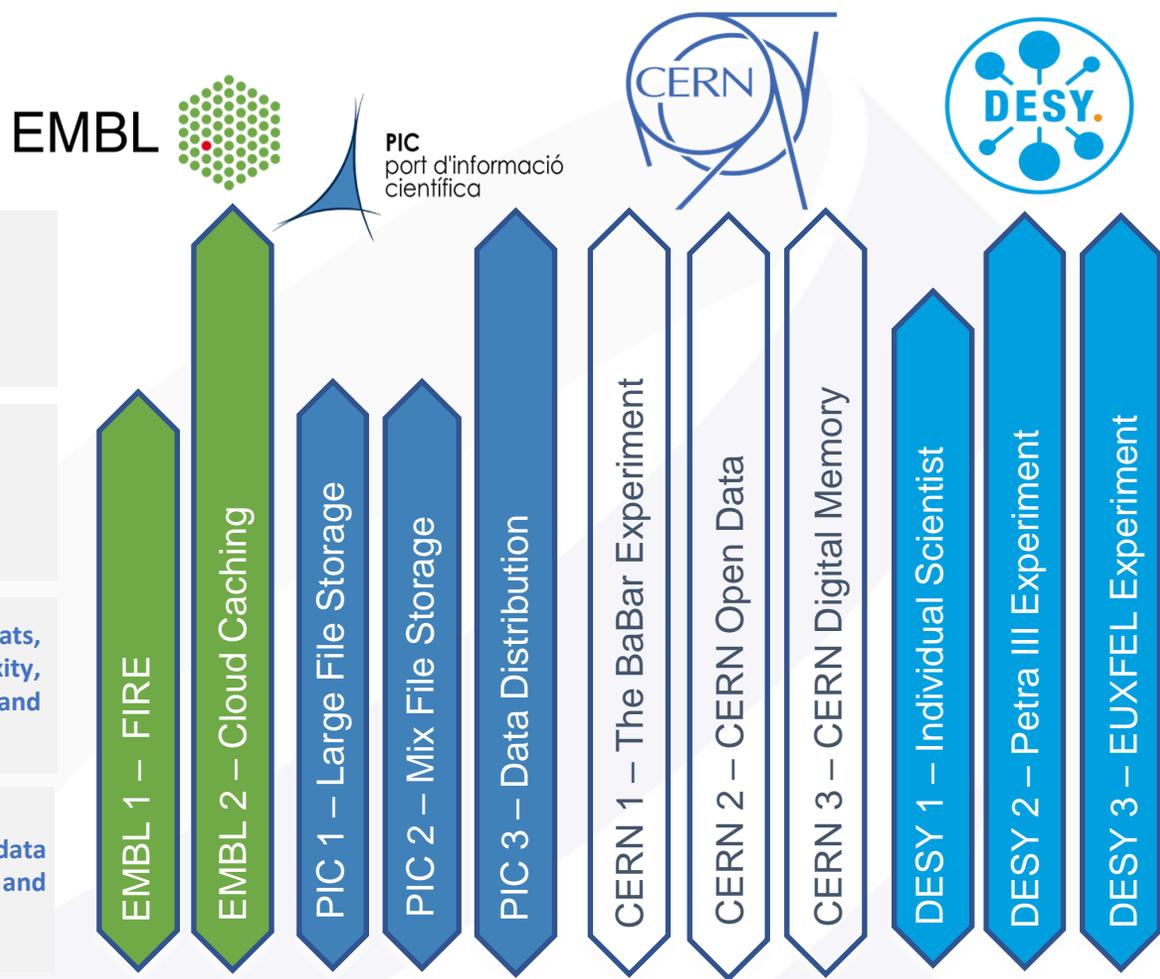
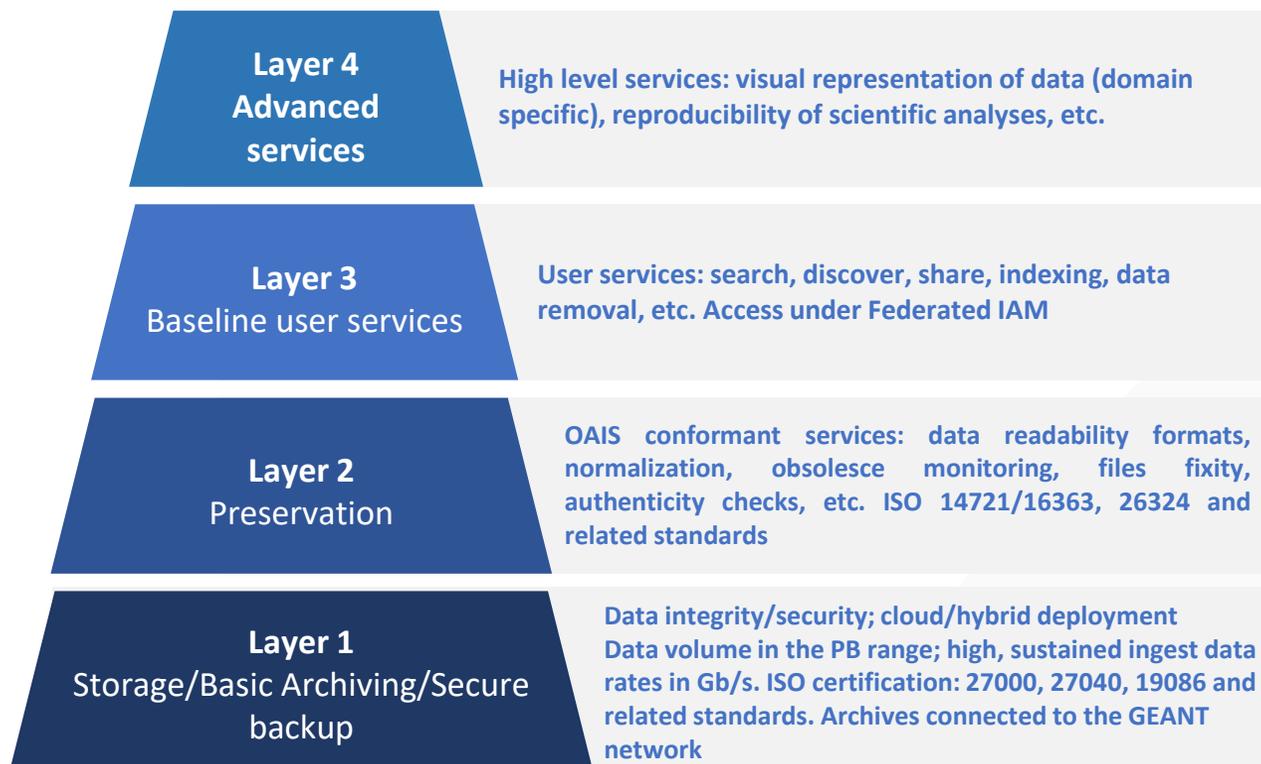
ARCHIVER is currently the only EOSC related H2020 project focusing on sustainable Archiving & LTDP services for PB scale datasets across multiple research domains and countries.

Progress Beyond the state of the art



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

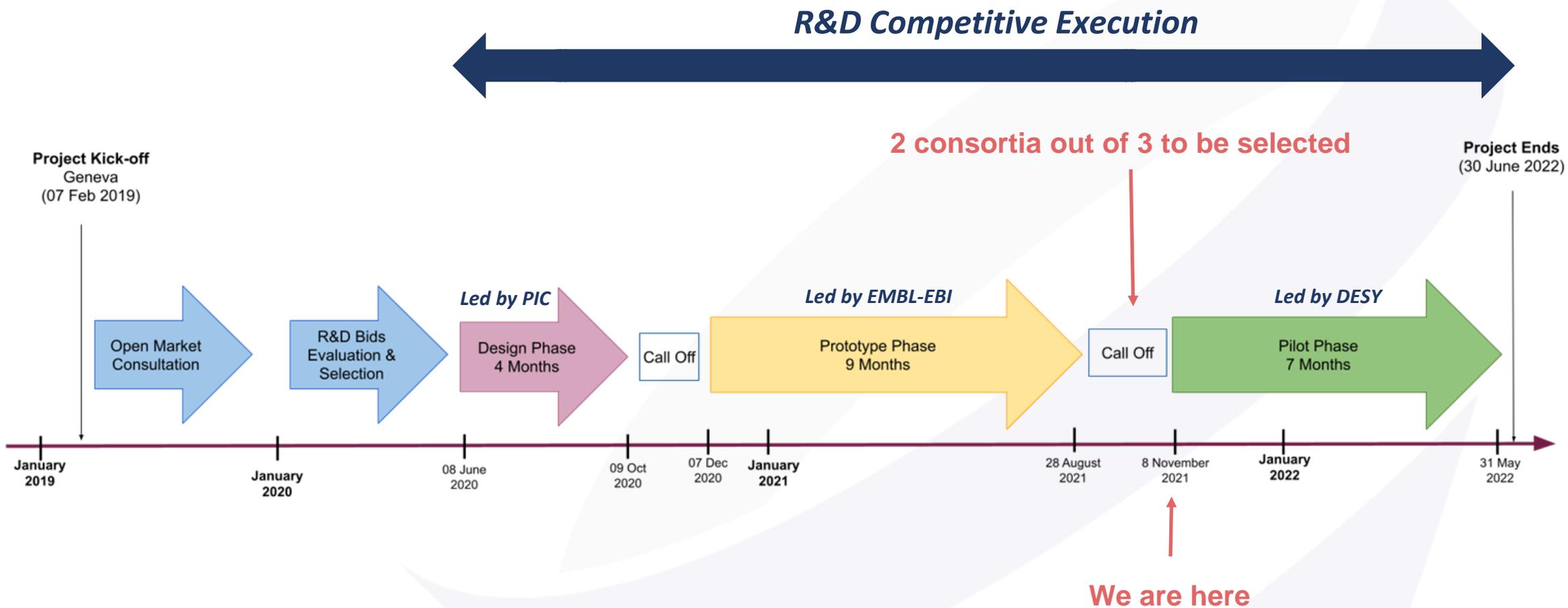
R&D Scope



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>

Competitive Process



Prototype Phase Consortia



arkivum

Bringing archived data to life



Google Cloud


libnova



UNIVERSITAT DE
BARCELONA



CSIC

CONSEJO SUPERIOR DE INVESTIGACIONES CIENTIFICAS



Giarretta
Associates



T-Systems



ONE DATA  GWWDG
Gesellschaft für wissenschaftliche
Datenverarbeitung mbH Göttingen

ICT infrastructure Environmental Metrics



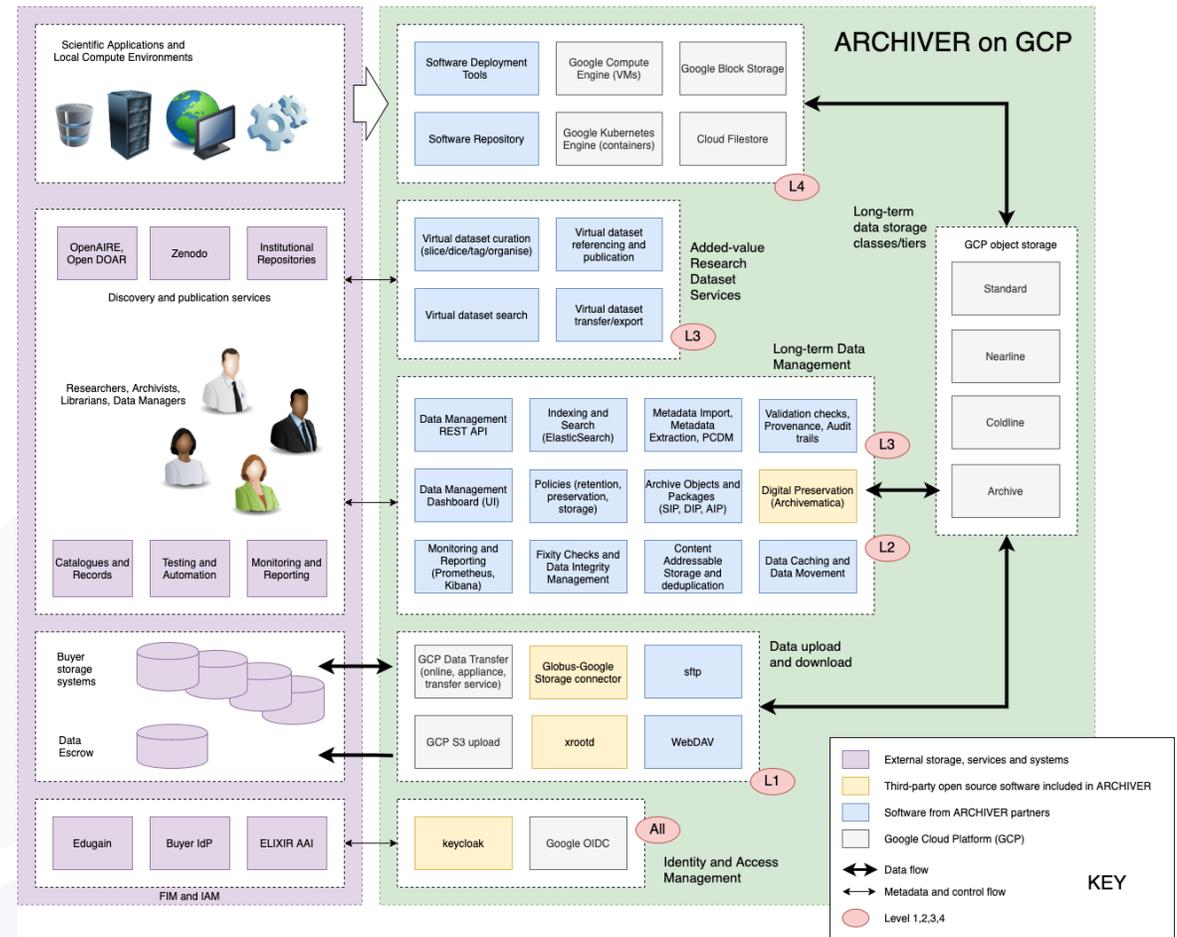
- **Power Use Effectiveness (PUE) ratio:** total amount of power consumed by a data center when compared with the total consumed by the server infrastructure
- **Carbon Emissions Intensity:** amount of carbon emitted in generating 1kWh of power - it can also be reflected in the **Renewable Energy Factor (REF)**
- **Server Utilization:** effective utilization of processing and storage capacity, depends on software applications and nature of research and organisation

$$\text{PUE} = \frac{\text{Total Facility Power}}{\text{IT Equipment Power}}$$



Environmental sustainable strategies: Arkivum

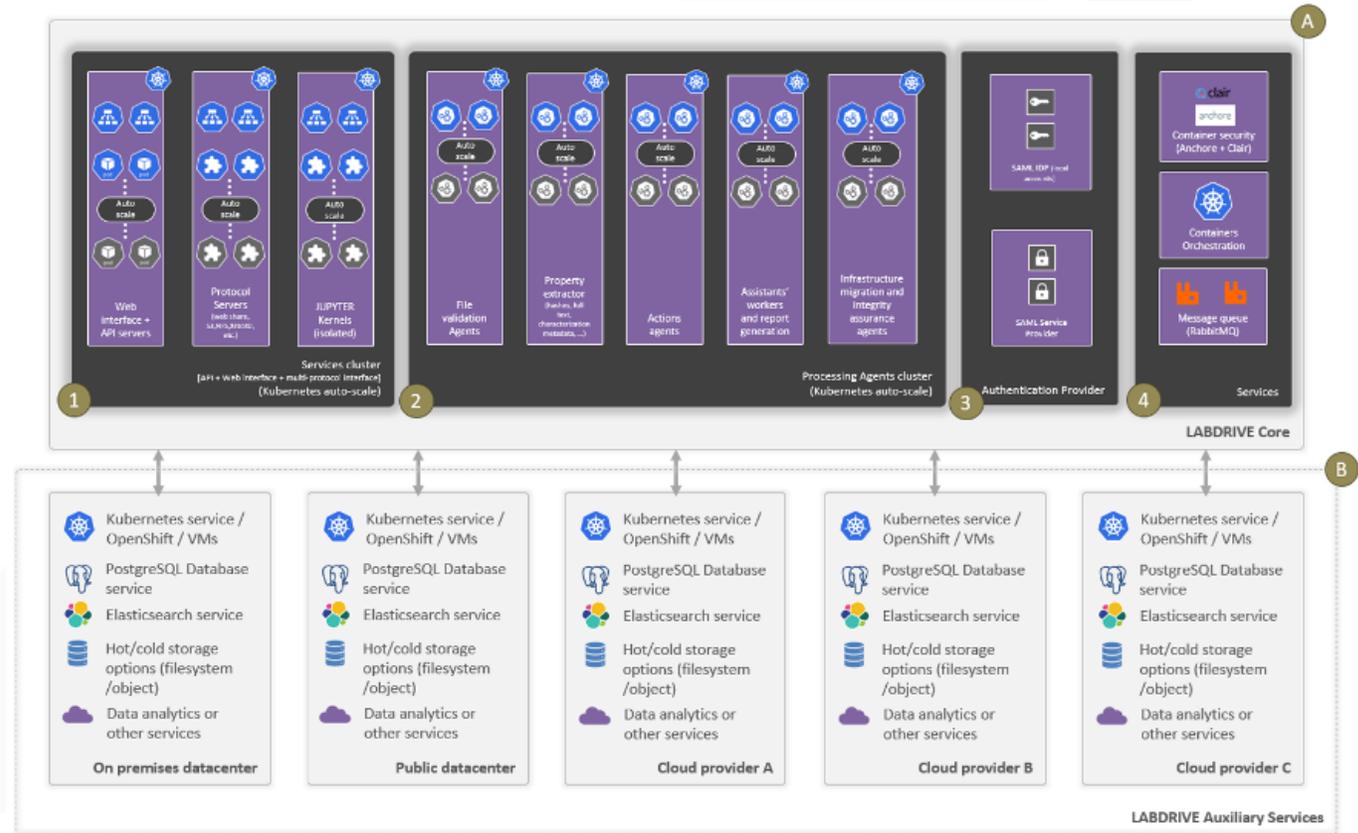
- Arkivum SaaS stack can be deployed on-premises or in a hybrid cloud configuration
- Google Cloud Platform (GCP) infrastructure carbon neutral since 2007, with multiple low carbon data centers in Europe: carbon free by 2030
- Overall architecture composed of micro-services to scale from 0 to multi-petabyte volumes of billions of objects to optimise use of resources.
- Based on Kubernetes containers: system auto-scales but does not consume resources when not under load



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

Environmental sustainable strategies: Libnova

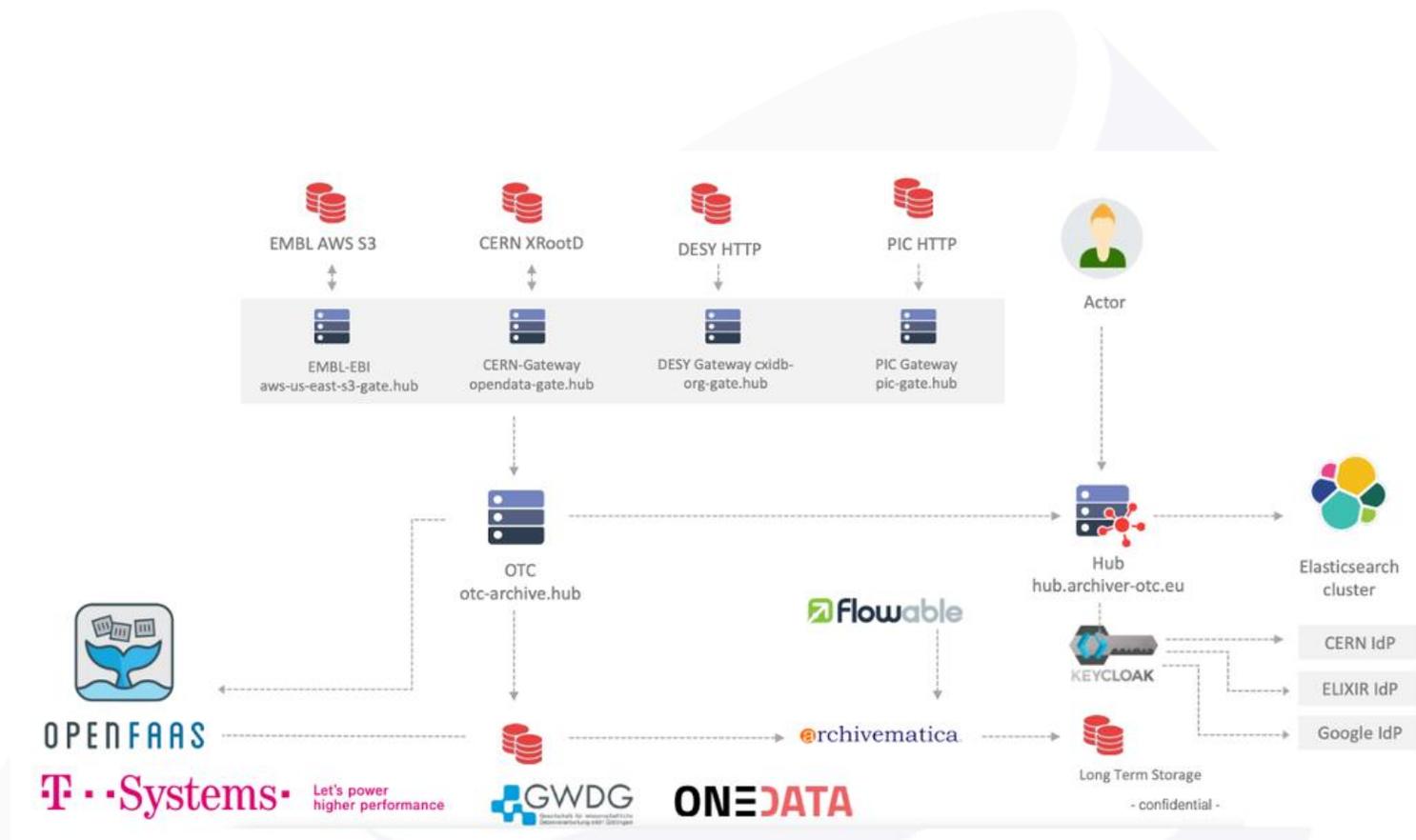
- Prototype based on LibSAFE SaaS
- Using infrastructure provided by AWS that aims to use only renewable power by 2025
- Software components running inside Kubernetes containers. Adjustable number of components/containers based on service demand to ensure full scalability and cost/environmental effectiveness.
- QoS optimisation of storage tiers considering carbon emissions among other factors.



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

Environmental sustainable strategies: T-Systems

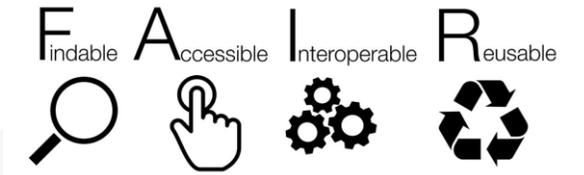
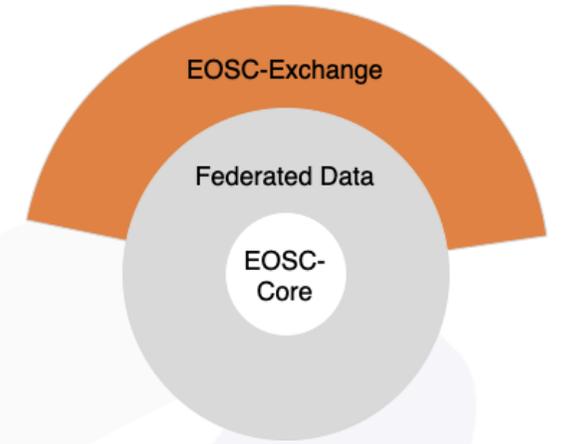
- Automated OSS architecture Onedata, OpenFaaS, Flowable & OTC infrastructure
- Gap analysis performed during the Design phase, to optimise Archivemata workflows for storage and network
- T-Systems OTC recently optimised the number of data centers per geographical area from 92 to 13 cutting operational carbon emissions by 50%.
- Kubernetes-based platform improving by 30% number of servers and ensuring portability to a different geographical areas (with lower carbon intensity electricity production)



Prototype architecture of the T-Systems consortium (image courtesy of the T-Systems 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**, with data processing strategies to minimize energy consumption and use of ICT infrastructure, fitting the EU GPP criteria.
- ARCHIVER pilot phase starts in November: exposing services to end-users and early adopters organisations to determine if they are suitable to their needs.





ARCHIVER

ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

Thank you! Questions?

If interested in knowing more about ARCHIVER, please register to the ARCHIVER Pilot Phase Kick-Off:

<https://archiver-project.eu/archiver-pilot-phase-kick-event>



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/UCCBlyLpUt-hWmQatqdlhlzw>