ARCHIVER

ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

Cloud-hosted, cost-effective long-term data management services

A solution for long-term data management and online access to address the challenges of how cloud hosted services can be used to store, manage, preserve and provide access to petabyte scale datasets.

The proposed solution

Arkivum and Google are designing and building a solution for long-term data management and online access to address the challenges of how cloud hosted services can be used to store, manage, preserve and provide access to petabyte scale datasets. The solution is designed to support research intensive organisations that are increasingly generating very large datasets from a range of sources, that need to be captured, ingested, digitally preserved and made accessible for others to use for the future. The solution will be deployed onto the Google Cloud Platform.

The solution focuses on archiving using **Trusted Digital Repository techniques** and ensuring data is Findable, Accessible, Interoperable and Reusable (i.e. adhering to the FAIR principles). The emphasis of the development is on **cost-effective archiving and preservation** that can meet the very large data volumes and high ingest rates from organisations such as CERN, DESY, EMBL-EBI and PIC. Understanding the **specific requirements of each different domain** and providing the appropriate solution constitutes the added value of Archivum solution.

The R&D potential

- Using **Google Cloud Platform (GCP) for a high speed ingest and access**, and also the ability to host and run scientific applications against the very large dimensions of datasets.
- Using **digital preservation and data archiving services to meet the requirements for OAIS** and to provide organisations with a hosted solution for operating their Trusted Digital Repositories.
- Ensuring that research data can be described, organised, sliced/diced, tagged and published in a flexible way that meets **FAIR principles**.
- **Total Cost of Service** models to optimize it against data volumes, access frequencies, data safety, data processing and retention periods.
- Use of open standards, open specifications, open-source and open APIs for **high-level portability, interoperability, exit strategies** & avoiding solution lock-in.
- Detailed **models and commercialisation plan**, including Service Level Agreements, User Support, Licensing, Service Configuration and Pricing for new commercial services based on ARCHIVER: this is a crucial element to be provided to the Buyers Group, the European Open Science Cloud (EOSC) and beyond.

Architecture overview

The Archivum solution, partnering with Google, will deliver the full ARCHIVER requirement stack. (see the 4 Layers here + link to the ARCHIVER R&D requirements above). The overall architecture is composed of micro-services to scale to multi-petabyte volumes of billions of objects, consisting of a service-oriented SaaS stack deployed on Google Cloud Platform (GCP) that addresses all four layers of ARCHIVER as described below. The solution can also be deployed on-premises or in a hybrid cloud configuration.

Next Page > Comparison between the levels of R&D before and after the introduction of ARCHIVER solutions

ARCHIVER

ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

Comparison between the levels of R&D before and after the introduction of ARCHIVER solutions

Date - January 2021

Baseline before ARCHIVER	ARCHIVER R&D
Storage/basic archiving/secure backup (Layer 1) Storage services deployments up to single PB scale.	Tens of PBs of scientific data volume with linear growth over the years. Demonstration of support of multiple tenancy with data and access isolated is required. Sustained data ingest rates capabilities from 1-10 GB/s.
Preservation (Layer 2) Preservation services support Preservation of files, folders, Content Management Packages (archives) and some data types such as Emails backup in pdf format.	Support of High level of redundancy, strong disaster recovery mechanisms, long-term planning for decades and active monitoring of data integrity in order to detect unwanted changes such as file corruption or loss. Best practices foreseen in CoreTrustSeal in terms of self-assessment. Support for handling unstructured and missing metadata. OSS components and vendor independent standards and interfaces (such as PREMIS, METS and Bagit) are preferred to allow implementation and demonstration during the Prototype and Pilot of exit strategies to prevent vendor lock-ins.
Baseline user services (Layer 3) User Services current support indexing, elastic search, deduplication, etc. for volumes of hundreds of TBs.	Support tools for search, look up or filter potential datasets rapidly, to access dataset metadata and decide on its relevance (e.g. citation purposes or reusing a dataset). Automated metadata indexing for several tens of PB content must be standard, aiming at maximum interoperability, including support for dataset filtering. Access and permission management against repositories and various collections supporting Federated Identity and Access Management. Fast information tagging and indexing for PB of data (easy and broader search, as a strategy to promote open data access).
Advanced Services (Layer 4) Several advanced services including retention and integrity capabilities of certain types of data over decades ensuring it is tamper-proof; Basic data re-use.	Prototype demonstrators of full reproducibility of services (initial examples are database services and/or software distribution services) on top of the resulting supported data archives. Ability to run additional scientific analyses independently of on-prem infrastructure in stages. For example: (i) Run scientific software distribution services on premise of the Buyers organisations relying on storage services provided externally by the commercial service provider, in hybrid mode, (ii) Run scientific software distribution services reproduced externally, relying on storage services provided by the commercial service providers.

Watch the video > Prototype Phase winners interview - Chris Sigley (ARKIVUM)
https://www.youtube.com/watch?v=xAscTZIe_Aw