ARCHIVER

ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

A standard-compliant data preservation solution

T-Systems International, GWDG and Onedata offer an OAIS-compliant data preservation solution, open, easy-to-use, extendable, cost- and energy-efficient, building on pre-existing and proven components for data preservation, data and workflow management.

The proposed solution

T-Systems' consortium provides research organisations and the European Science community with an OAIS-compliant data preservation solution, that is open, easy-to-use, extendable, cost- and energy-efficient. The solution follows a **full open-source and cloud-agnostic approach**, building on pre-existing and proven components for data preservation, data and workflow management. The core components include Archivematica, Onedata and Flowable, and have been selected based on functionality, integration, cloud-adoption, maturity, size of user community and relations with EOSC. The modular approach is supported by a large set of APIs that will enable users to extend and integrate the components with other preferred services.

The R&D potential

The R&D focuses on new innovative functionality for baseline and advanced data preservation services, including Petabyte-scale storage options, compliance with OAIS, PREMIS, METS and BagIT standards and new innovate functions for distributed data and workflow management, search and discovery, data representation and scientific analysis. The objective is an **integrated service offer for end-users**, integrators and cloud providers for deployments on local and public cloud infrastructures. T-Systems operates the services as part of its Open Telekom Cloud portfolio, a leading European public cloud service based on OpenStack. GWDG extends its portfolio with the service offered for its established public and academic community.

Architecture overview

A modern Open-Source architecture based on Onedata, Archivematica, OpenFaaS, Flowable and OTC computing and storage services integrated into a single Kubernetes-based platform with BPMN (Business Process Model and Notation)- based workflow Management.

Next Page > Comparison between current baseline and R&D being performed

ARCHIVER

ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

Comparison between current baseline and R&D being performed

Date - January 2021

Baseline before ARCHIVER	ARCHIVER R&D
Storage/basic archiving/secure backup (Layer 1) A few solutions exist providing high performance scalable cloud storage. Even fewer when considering multi-cloud transparent data access and many closed sources, making it difficult to customize or add domain-specific features. More solutions exist which came from the domain of scientific data processing communities. iRODS36 is a data management system, in which the users have to manage the data location by themselves.	Data management relying on the features of Onedata, a distributed virtual filesystem platform allowing seamless integration of a wide range of legacy storage resources as transfer sources, QoS based replica placement, metadata management and open data support.
Preservation (Layer 2) **Baseline user services (Layer 3)** **Advanced Services (Layer 4)** Only a few considered end-to-end preservation services are open-source, such as Archivematica. A detailed gap analysis shows support for open standards such as PREMIS, METS or BagIt, but as well considerable limitations and drawbacks including focus on text documents, images or other small files rather than large opaque computational data sets. Interfaces are optimized for manual workflows rather than continuous automatic ingestion pipelines, limited capabilities for scaling the metadata extraction and other pre-ingestion microservices. Archivematica shows limitations in scalability and is in need of a general optimisation of the supported workflows also in terms of cost effectiveness: currently, Archivematica data ingests need 3 times more storage for archive preparation than the volume of the data ingested.	The proposed architecture has chosen Archivematica as the basis for the data archiving and preservation functionality, due to its fully open-source license, standards compliance and active development community. R&D plan was conceived to extend the data archiving and preservation by capabilities using a Onedata plugin to address the identified limitations. Integrated architecture proposal with R&D effort in improving the Archivematica workflow with OSS packages in areas such as scalability, data workflow management and data ingest cost effectiveness: · Support of arbitrary size standards compliant archive-packages in a cost- effective manner (Layer 1): no large storage required anymore for archive preparation. · Adds OpenFaaS serverless functions for scalable metadata extraction with containers (Kubernetes), metadata-based data discovery and support for long running asynchronous tasks and autoscaling features. (Layer 2/3) · Adds Flowable for automated coordination of archiving workflows; (Layer 2/3) · Onedata plugin to Archivematica, allows automatic archive ingestion from Onedata spaces; (Layer 3/4) · Onedata provides open data functionality for Metadata management. Data discovery, OAI- PMH interface and DOI and PID minting. (Layer 3/4) · Supports Kubernetes, as the main deployment framework for the proposed platform for automated independent deployment at scale (Layer 4). · Kubernetes allows flexible platform configuration through Helm Charts, extensive modularization through Pods and allows easy file system sharing between containers (Layer 4).

Watch the video > Prototype Phase winners interview - Jurry de la Mer (T-SYSTEMS)
https://www.youtube.com/watch?v=2-tdDGcrp_s