IJDC | *Brief Report*

Towards environmentally sustainable long-term digital preservation

Ignacio Peluaga
CERN

João Fernandes
CERN

Shreyasvi Natraj
CERN

Abstract

ARCHIVER and Pre-Commercial Procurement funding has enabled small to medium enterprises (SMEs) to innovate and deliver new services for EOSC. Within the framework of the ARCHIVER pre-commercial procurement tender, between December 2020 and August 2021, three commercial consortia competed to deliver innovative, prototype solutions for long-term data preservation. Two of them were selected to continue with the pilot phase and deliver research-ready solutions for long-term data preservation of research data, therefore filling a gap in the current European Open Science panorama.

Digital preservation relies on technological infrastructure (information and communication technology, ICT) that can have environmental impacts. While altering technology usage can reduce the impact of digital preservation practices, this alone is not a strategy for sustainable practice. Moving toward environmentally sustainable digital preservation requires critically examining the motivations and assumptions that shape current practice. The use of scalable cloud infrastructures can reduce the environmental impacts of long-term data preservation solutions.

Submitted 13th March 2022 ~ Accepted 22nd April 2022

Correspondence should be addressed to Ignacio Peluaga, CERN. Email: ignacio.peluaga.lozada@cern.ch

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Copyright rests with the authors. This work is released under a Creative Commons Attribution License, version 4.0. For details please see https://creativecommons.org/licenses/by/4.0/

http://dx.doi.org/10.2218/ijdc.v16i1.848
DOI: 10.2218/ijdc.v16i1.848

Introduction

Launched in January 2019 and with a total budget of €4.8M, co-funded by the European Commission under grant number 824516, the ARCHIVER project focused on archiving and data preservation services. It had a duration of 42 months and followed the Pre-Commercial Procurement format (2007). The project was led by CERN with three other organisations forming the Buyers Group (BG): CERN[1] in Switzerland, DESY[2] in Germany, EMBL-EBI[3] in the UK and PIC[4] in Spain. These publicly funded research organisations committed funds, use cases and testing efforts. Two other partners joined the consortium to bring expertise in requirements assessment and promotion activities: Addestino and TrustIT.

The BG analysed the state of the art prior to starting the procurement and identified the domains to be pursued throughout the project, such as PB scale volumes, involvement of European providers and application of best practices in the preservation sector (for example, the FAIR principles (Wilkinson et al., (2016), repository trustworthy certification (CoreTrustSeal Standards and Certification Board. (2019), and digital preservation capacity (Digital Preservation Coalition (2021)) among others. All of this while promoting open software and standards and considering the environmental impact.

R&D Execution

ARCHIVER followed an implementation based on three phases with multiple competing consortia. The activity during each phase produced the results to be taken into account in the selection process that allowed a contractor to proceed to the subsequent phase:

- **Phase 1** - Solution Design: Contractors developed designs including architecture and technical components.
- **Phase 2** - Prototype Development: Contractors selected from the Design Phase built prototypes based on the designed solutions.
- **Phase 3** – Pilot Deployment: Contractors selected from the Prototype Phase deployed expanded prototype services for testing with the identified use cases.

To prepare the R&D execution, the BG provided a set of use cases[5] (see Figure 1.) that had demanding archiving and preservation requirements in combination with significant volume sizes, sustained high ingest rates and long retention periods.

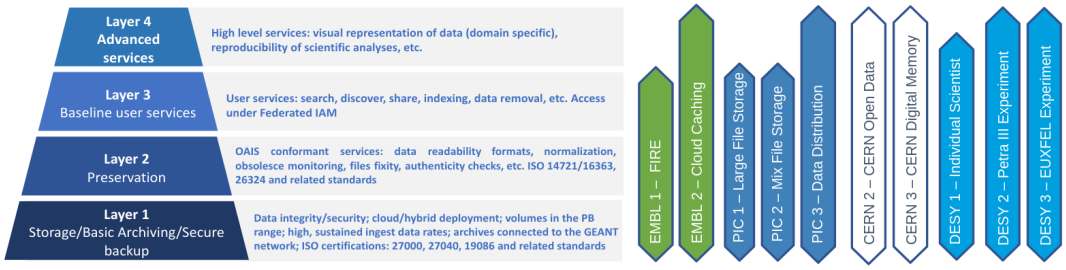

Figure 1. Mapping of use cases to the R&D challenge layers.

[1] Organisation européenne pour la recherche nucléaire: https://www.cern.ch/

[2] Deutsches Elektronen-Synchrotron: https://www.desy.de/

[3] EMBL's European Bioinformatics Institute: https://www.ebi.ac.uk/

[4] Port d'Informació Científica: https://www.pic.es/

[5] Deployment Scenario Technical Summaries: https://www.archiver-project.eu/deployment-scenarios

Timeline

Before the R&D competitive execution, there was a preparation period which included an open market consultation and the tender. A total of 15 bids to the tender were received. In June 2020 the competitive execution started, which lasted two years and included the three phases explained in the previous section, as seen in Figure 2.

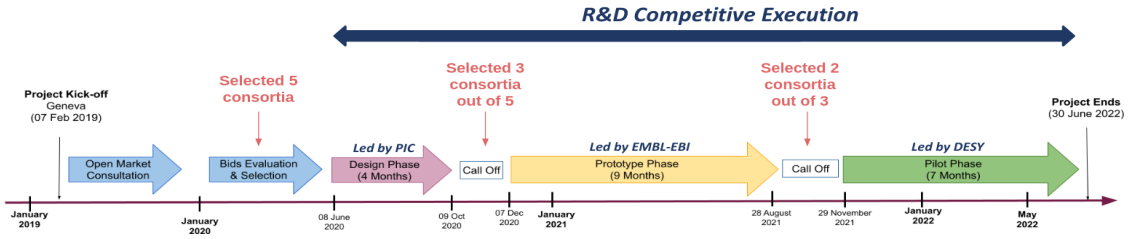

Figure 2. Timeline of the ARCHIVER project.

Selected Consortia

Five out of the 15 bids were initially selected to participate in the R&D competitive process. Two were selected for the final pilot phase: Arkivum and LIBNOVA.

Arkivum SaaS

The Arkivum solution runs on Kubernetes on the Google Cloud Platform (GCP). It takes advantage of auto-scaling which reduces costs and carbon emissions as resources are only used when they are needed (see Figure 3).

Figure 3. Example of cluster scaling on the Arkivum solution.

Additionally, users can select different storage options, which have prices linked to the access frequency: for example, infrequently accessed storage is cheaper.

GCP's infrastructure has been carbon neutral since 2007, with multiple low carbon data centres across Europe. GCP plans to become carbon free by 2030.[6]

[6] GCP Cloud Sustainability: https://cloud.google.com/sustainability

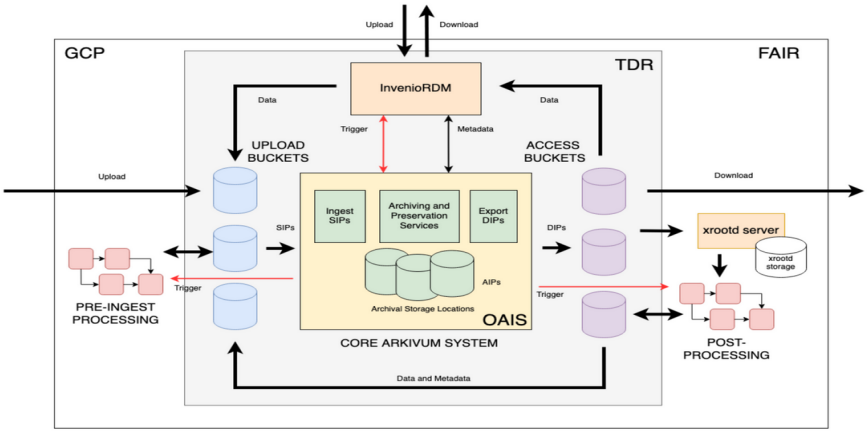

Figure 4. Diagram of the Arkivum solution's architecture.

LIBNOVA SaaS

In the context of the ARCHIVER Project, LIBNOVA has developed the LABDRIVE platform (Giaretta & Redondo, 2022). It is deployed on Amazon Web Services (AWS) and runs on Kubernetes, enabling auto-scaling to reduce both cost and carbon footprint. LABDRIVE was internally tested to receive 50 million files and 1PB of data in less than 24 hours, scaling itself to more than 6,500 Kubernetes pods to process the workload.[7] In terms of storage, LABDRIVE gives the user the option to select between different storage classes: for example, S3 Glacier has a lower price than standard S3. Similarly, to GCP, AWS is also committed to reducing the impact its cloud platform has on the environment and aims at using renewable-only energy by 2025.[8]

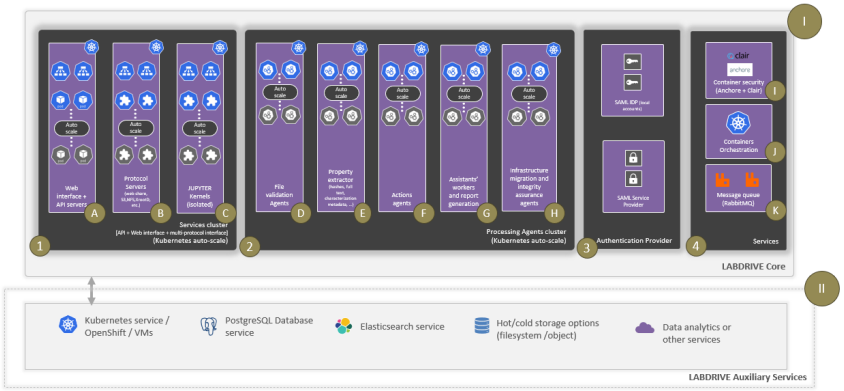

Figure 5. LIBNOVA's *LABDRIVE* system architecture.

FAIR in ARCHIVER

Following the FAIR principles can contribute to reducing the carbon footprint, given that if data can be found, accessed and reproduced, it's likely duplication will be avoided. In the context of the project, it was a requirement that the platforms developed by the contractors could implement the FAIR principles. To assess the degree of FAIRness of the produced repositories, a process of three steps was put in place:

[7] LABDRIVE System Architecture: https://docs.libnova.com/labdrive/concepts/architecture#system-architecture
[8] AWS Energy Transition: https://aws.amazon.com/energy/sustainability/

1. **Datasets ingestion -** Ingest data from the BG with metadata in DataCite or DublinCore format.

2. **Automated FAIR assessment -** Performed with FAIRsFAIR developed tool F-UJI (Devaraju & Huber, 2021).

3. **Manual FAIR assessment -** The automated assessment was complemented with a manual approach, with the support of reports provided by the contractors.

Conclusions

The R&D challenge of digital archiving goes beyond simply storing data. It is crucial to keep intellectual control of the data and associated products for decades, making research outputs reusable. ARCHIVER acted as a vehicle to commoditise archiving and preservation in the research domains and it has promoted 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-premises installation is possible as well to reuse procuring organisations' existing infrastructure.
- The combination of automation, microservices, serverless computing and cloud IaaS enable more efficient use of resources.
- The selection of cloud location/provider using environmentally friendly infrastructures can improve environmental sustainability.
- Make smarter use of storage e.g., deep/cold/infrequent access archive, small footprint access copies.
- Enable FAIR to increase re-use of results and avoid consuming energy to reproduce results.

There is strong evidence of the suitability of the resulting services with high levels of innovation in aspects such as scalability, robustness, and flexibility. The ARCHIVER Project was nominated for the International Council on Archives (ICA) Award for Collaboration and Cooperation.[9]

[9] Digital Preservation Awards 2022 Finalists Announced: International Council on Archives Award for Collaboration & Co-operation: https://www.dpconline.org/news/dpa2022-cc-finalists-announced

References

Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions (2007) Pre-commercial Procurement: Driving innovation to ensure sustainable high quality public services in Europe. https://op.europa.eu/en/publication-detail/-/publication/9d2e47d1-b0f3-11ec-83e1-01aa75ed71a1

CoreTrustSeal Standards and Certification Board. (2019). CoreTrustSeal Trustworthy Data Repositories Requirements: Extended Guidance 2020–2022 (v02.00-2020-2022). Zenodo. https://doi.org/10.5281/zenodo.3632533

Digital Preservation Coalition Rapid Assessment Model (2021) DPC RAM. http://doi.org/10.7207/dpcram21-02

Devaraju, A. and Huber, R. (2021). An automated solution for measuring the progress toward FAIR research data. Patterns, vol 2(11), https://doi.org/10.1016/j.patter.2021.100370

Giaretta, David Leslie, & Redondo, Teo. (2022). Building LABDRIVE, a Petabyte scale, OAIS/ISO 16363 conformant, environmentally sustainable archive, tested by large scientific organisations to preserve their raw and processed data, software and documents. https://doi.org/10.5281/zenodo.6636295

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 https://doi.org/10.1038/sdata.2016.18