ARCHIVER

Sustainable Preservation of Scientific Data

|  |  |  |
| --- | --- | --- |
| **Matthew Addis** | **Teo Redondo** | **João Fernandes** |
| *Arkivum Ltd*  *UK*  *matthew.addis@arkivum.com*  *0000-0002-3837-2526* | *LIBNOVA*  *Spain*  *teo.redondo@libnova.com*  *0000-0001-6465-7771* | *CERN*  *Switzerland*  *joao.fernandes@cern.ch*  *0000-0002-0445-7038* |
|  |  |  |

**Abstract – The ARCHIVER project (Archiving and Preservation for Research Environments) has spent over 3 years designing, prototyping and piloting innovative new services for the Long Term Digital Preservation (LTDP) of scientific datasets. During the project, multiple Data intensive organizations representing several research domains (CERN, DESY, PIC and EMBL-EBI) have worked closely and collaboratively with suppliers (Arkivum and LIBNOVA) on the research and development of new services and solutions for scientific data preservation relevant for the European Open Science Cloud (EOSC). This panel session will see the ARCHIVER project participants discuss and share the experience and lessons learned during the project. Topics will include: the benefits of a collaborative approach between end-users and commercial suppliers; the challenges that were addressed along the way and the solutions that were created; and what still needs to be done in order to realize the project vision of sustainable digital preservation services for the whole scientific community that address the needs of organizations both large and small. iPRES 2022 comes just two months after the end of the final Pilot phase of the ARCHIVER project which makes it an ideal time for the ARCHIVER participants to share their insights and experiences with the wider LTDP community.**

**Keywords – Digital Preservation, Scientific Data, Trusted Digital Repository, Sustainability, Scalability**

**Conference Topics - Environment; Innovation.**

1. Introduction

With a procurement budget of 3.4 million euros, the ARCHIVER project [1] has used a Pre-Commercial Procurement (PCP) approach to competitively procure R&D services from a range of vendors in order to create new services and solutions for LTDP of scientific datasets. The three stages of the project cover design, prototyping and pilots and have taken place from Jan 2019 - Jun 2022.

ARCHIVER is driven by the needs of a diverse range of stakeholders including CERN, who operate the Large Hadron Collider near Geneva, DESY (the Deutsches Elektronen-Synchrotron, based in Hamburg and Berlin), the EMBL-EBI (European Bioinformatics Institute, based in Cambridge), and PIC (Port d’Informació Científica, situated near Barcelona).

The importance and benefits of making scientific data open and reusable according to FAIR principles [2] (Findable Accessible Interoperable Reusable) has been clear for some time now. Practical advice and guidelines are now available from initiatives such as the FAIRsFAIR [3] project. However, there are still major gaps [5] when it comes to long-term accessibility and usability of research data, for example as discussed in the FAIR Forever report [4] from the Digital Preservation Coalition. These gaps put research data at long-term risk, they prevent the construction and operation of sustainable Trusted Digital Repositories [6], and they affect organizations both large and small who are tasked with being custodians of valuable research data resources. This is what the ARCHIVER project sets out to solve.

The aim of ARCHIVER is to achieve substantially improved archiving and digital preservation, not just for petabyte-scale data-intensive research, but also for the Long Tail of Science (LToS) [7]. To support the requirements of European scientists, ARCHIVER provided R&D funding to European SME sector specialist to stimulate new end-to-end archival and preservation services for the vast and ever-growing datasets generated by world-leading research institutions. Reflecting the move toward large-scale collaborative research supported by cloud infrastructures, ARCHIVER embraces and tackles issues such as scalable and interoperable LTDP services in the cloud, new business and commercial models for archiving, accessing and reusing large datasets, and crucially how to do this in a way that is both economically and environmental sustainability.

This panel will discuss the ARCHIVER project, the results that have been achieved, the approaches that were taken, what worked and what didn’t, what still needs to be done to make LTDP a reality for the scientific data community, and how the ARCHIVER experience and approach could be translated into other domains and markets. The aim of the panel is to follow the ARCHIVER spirit of being open, honest and transparent and to share our experience and thoughts with the community as a whole.

# Panel Discussion Topics

The panel will discuss a range of topics and questions that include:

* What worked well in ARCHIVER (and what needs improvement) when the suppliers and end-users collaborated together using the project’s pre-competitive R&D approach? Do you think this could be a good template for other sectors?
* What were the main challenges in ARCHIVER and how were they overcome? Were they organisational, technological, economic or a mix of all these things?
* What are the three most important features of each of the ARCHIVER resulting solutions? How did the end-users articulate their priorities, and how did the suppliers go about building their solutions?
* How can LTDP services support organisations who produce research data from the large scale through to the long tail, for example in the context of initiatives such as the European Open Science Cloud? Are there economies of scale, is there a one size fits all, and how can the LToS benefit from the services developed in ARCHIVER that are primarily for large organisations?
* Is it possible to preserve and provide access to huge volumes of data in a way that is environmentally sustainable? Does the cloud help, or does it make it harder? What does the carbon footprint of LTDP look like in practice?
* How does digital preservation fit into making data FAIR Forever? How do digital preservation standards and good practices help organizations build trusted repositories?
* The value of scientific data is often in its reuse, for example re-running computations and applications against archived data. How do the ARCHIVER resulting services address the need to both preserve data and at the same time support active access and reuse?
* ARCHIVER has done a lot of work on technological solutions, but what about the economics and business models? What commercialisation approaches do the ARCHIVER team foresee for the resulting services developed in the context of the project?
* Research data lives for longer than any vendor, system or technology. How do the ARCHIVER resulting LTDP services prevent vendor lock-in and encourage portability and interoperability, yet at the same time make it attractive for new commercial services to enter the market? Are these in conflict with each other?
* Do you think that the lessons learnt and the solutions developed in ARCHIVER are transferable from ARCHIVER to other disciplines and domains? What would be your number one recommendation?

REFERENCES

1. [www.archiver-project.eu](http://www.archiver-project.eu)
2. Findable Accessible Interoperable Reusable (FAIR).<https://www.go-fair.org/fair-principles/>
3. <https://www.fairsfair.eu/>
4. A. Curry, W. Killbride. “FAIR Forever? Long Term Data Preservation Roles and Responsibilities, Final Report”. Feb 2021. [https://zenodo.org/record/4574234](https://zenodo.org/record/4574234#.YD5_F2j7SUk)
5. J. Fernandes et al. “ARCHIVER D2.1- State of the Art, Community Requirements and OMC Results “. Jan 2020. [https://zenodo.org/record/3618215](https://zenodo.org/record/3618215#.YIl2SubTWJR)
6. D Lin et al. “The TRUST Principles for digital repositories”. Scientific Data. 7, Article 144. May 2020.
7. M. Devouassoux, B. Jones, J. Fernandes. “Long-Tail-of-Science's Requirements for Commodity Cloud Services in Europe” Oct 2019. https://zenodo.org/record/3564668