ARCHIVER

ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

Deliverable Title: 2.1 - State of the Art, Community Requirements and OMC Results Report

Partner Responsible: CERN, ADDESTINO

Work Package: WP2

Submission Due Date: M6

Actual Submission Date: 14/11/2019

Distribution: Public

Nature: Report

Abstract: This report highlights the results of the ARCHIVER Pre-Commercial Procurement (PCP) Open Market Consultation (OMC) process. It includes a State of the Art analysis of the digital archiving and preservation solutions available both on the supply and the demand side and an assessment of the community requirements stemming from the landscaping activities carried out during the OMC.

Document Information Summary	
Deliverable number:	D2.1
Deliverable title: Editor:	State of the Art, Community Requirements and OMC results Report Marion Devouassoux (CERN), João Fernandes (CERN)
Contributing Authors:	Dominique Buyse (Addestino), Ruben Van Calenberg (Addestino), Vaggelis Motsenitalis (CERN), Jakub Urban (CERN)
Reviewer(s):	Marion Devouassoux (CERN), João Fernandes (CERN), Bob Jones (CERN)
Work Package no.:	WP2
Work Package Title:	Tender Preparation
Work Package Leader:	Addestino
Work Package Participants:	CERN, EMBL-EBI, DESY, PIC and TRUST-IT Services
Distribution:	Public
Nature:	Report
Version/Revision:	V 1.1
Draft/Final:	Resubmitted following EC review
Keywords:	OMC, Community Landscaping, State of the Art

Disclaimer

The ARCHIVER project with Grant Agreement number 824516 is a Pre-Commercial Procurement Action funded by the EU Framework Programme for Research and Innovation Horizon 2020.
This document contains information on the ARCHIVER core activities, findings, and outcomes, and it may also contain contributions from distinguished experts who contribute to ARCHIVER. Any reference to content in this document should clearly indicate the authors, source, organisation, and publication date. This document has been produced with co-funding from the European Commission. The content of this publication is the sole responsibility of the ARCHIVER consortium and cannot be considered to reflect the views of the European Commission.

Grant Agreement Number: 824516
Start Date: 01 January 2019
Duration: 36 Months

Document History

Issue	Date	Description	Author/Partner
V0.1	21/06/2019	First draft version available for the project office	Marion Devouassoux (CERN), Vaggelis Motesnitalis (CERN)
V0.2	24/06/2019	Edits and comments	Vaggelis Motesnitalis (CERN) Marion Devouassoux (CERN), João Fernandes (CERN), Joshua Davison (CERN)
V0.3	28/06/2019	Edits and comments	João Fernandes (CERN), Bob Jones (CERN)
V0.4	11/07/2019	Major update	João Fernandes (CERN)
V0.5	17/07/2019	Substantial edits across	João Fernandes (CERN), Bob Jones (CERN), Marion Devouassoux (CERN)

Document Approval

Issue	Date	Name
V0.1	22/06/2019	First draft
V0.2	24/06/2019	Second draft
V0.3	28/06/2019	Updated version shared with the ARCHIVER PO
V0.4	11/07/2019	Updated version shared with the ARCHIVER PO
V0.5	30/07/2019	Updated version shared with the ARCHIVER Consortium
V1.0	15/08/2019	Final version to be submitted to the EC
V1.1	10/10/2019	Resubmitted following EC review

Executive Summary

Many research projects are currently struggling to preserve their data and associated products (metadata, software, documentation, etc.), as the current archiving and preservation capabilities are inadequate in terms of scale and fall below expectations for a number of communities, while data stewardship costs are frequently underestimated during the planning phase.
The goal of ARCHIVER is to fulfil these data management promises in a multi-disciplinary environment, allowing each research group to retain stewardship of their data whilst leveraging best practices, standards and economies of scale.

The ARCHIVER consortium started by performing an analysis to identify the current gaps in preservation services offered in the public sector, taking lessons learned from past initiatives and consolidating future opportunities for the ARCHIVER resulting services.
A key finding during the state of the art analysis is that ARCHIVER will benefit from the know-how the research community accumulated through the development of models for successful curation and preservation of data and associated assets, such as documentation and software, across all stages of the curation cycle. ARCHIVER will put these models into practice for multiple scientific disciplines in order to simplify data management, making costs more predictable and helping to reduce fragmentation in data stewardship practices across scientific domains.
An additional result of the analysis is the fact that the current context of the European Open Science Cloud (EOSC) provides an unprecedented opportunity to make the resulting services available to an estimated 1.7 million researchers in Europe. Services to make data Findable, Accessible, Interoperable and Reusable (FAIR), to store it and ensure long-term preservation will form the core of the EOSC. ARCHIVER will thus contribute with a set of aligned services for scientific data management, following best practices with clear criteria defined for the selection of high quality trustworthy repositories, serving FAIR data to scientists and consequently increasing data reuse fostering the development of science.

In order to stimulate an open dialogue that broaches the views of the market about the intended R&D scope, ARCHIVER organised an Open Market Consultation (OMC) with a series of activities for both demand and supply sides. The dialogue within the community, end-users and

potential tenderers of the project took place between April and June 2019 as announced in a Prior Information Notice[1] published in the Official Journal of the European Union. These activities pursued a three-fold objective:

- Inform the research community, supply-side and the demand-side of the project and the upcoming call for tender
- Assess the requirements for digital preservation and archiving services of the research community at large and specifically in the context of the EOSC
- Improve the mutual understanding of the R&D challenges across procurer organisations, future early adopters organisations and industry in order to verify the innovation potential and feasibility of the project.

The OMC process brought together over 35 companies, the majority of which are Small and Medium Enterprises (SMEs), as well as many public organisations in need of innovative data archiving and preservation services.
Feedback from the participating supply-side companies in the OMC events has been collected, and the dialogue progressively focussed on the Buyers Group deployments scenarios requirements for the resulting ARCHIVER services.

The OMC activities resulted in the following outputs:

- Assessment of the current needs from the research community for archiving and preservation services at large and in the context of the EOSC
- Prioritization of the main ARCHIVER challenges
- Estimation of value, risk and effort of the R&D to be performed in ARCHIVER
- Gathering of a number of lessons learned for future PCP projects

More broadly, ARCHIVER analysed the current state-of-the-art in order to establish a technological baseline state of play, assessing the current solutions available in the market with an analysis of the progress from the current state-of-the-art and advancement to be demonstrated by the ARCHIVER R&D activities.

[1] ARCHIVER Prior Information Notice:
https://ted.europa.eu/TED/notice/udl?uri=TED:NOTICE:87071-2019:TEXT:EN:HTML&src=0&ticket=ST-13546038-zzOsx8dRAbr0FP3r9xzt3AU6BwKF0ikzmaysCZ3lLrjuh9JMXyEfinlSbBWSbkG7nVMM8Q7175Fqeyl6q12Wwm8-rS0vSrmBGYCqxcQHPaJ3L4-SEkdHJzdaDPWnhuuEnm8R1K1eE2su2bUD3gPyvmidzp0

These findings are reflected in the PCP Contract Notice, specifically in the selection and award criteria of the functional specifications, in order to ensure that the selected bids are capable of meeting ARCHIVER's R&D challenge.

Table of Contents

Acknowledgments

The ARCHIVER consortium would like to thank the following organisations for their involvement in the Open Market Consultation and landscaping activities:

- The CERN Industrial Liaison Officers (ILOs) for the help provided in disseminating the OMC events to potential tenderers across their countries.
- The Catalonia Trade & Investment Agency (ACCIÓ)[2] and Europe Entreprise Network (EEN)[3] for promoting the Open Market Consultation Events
- The Crowdhelix[4] team for making the Crowdhelix platform available at no cost to companies looking for partners to build consortia in order to bid for the tender.
- The Digital Preservation Coalition (DPC)[5] for its collaboration with the project providing expertise including a tailor-made webinar training on Digital Preservation and Open Archival Information System (OAIS)[6] during the project's Open Market Consultation

1. Introduction

This report highlights the results of the ARCHIVER Pre-Commercial Procurement (PCP) Open Market Consultation (OMC) process. The document is structured in 4 main sections:

- State of the Art Analysis:
  It includes an assessment of the current and past activities of a number of research communities actively involved in data preservation including some with experience

[2] http://catalonia.com/about-accio.jsp
[3] https://een.ec.europa.eu/
[4] https://www.crowdhelix.com/
[5] https://www.dpconline.org/
[6] https://www.archiver-project.eu/training-digital-preservation-and-open-archival-information-system-oais

spanning decades; It includes as well an analysis of lessons learned from similar R&D activities.

- Open Market Consultation Roadmap:
  It provides an overview of all the activities ARCHIVER has undertaken, both from the demand and supply-sides, and community landscaping efforts.

- Main Outcomes of the OMC process for demand and supply sides.

- Technological baseline state of play:
  It includes an assessment of the current solutions available in the market with an analysis of the progress of the current state-of-the-art and advancement to be demonstrated by the ARCHIVER R&D activities.

- Conclusions and lessons learned for future PCPs

The OMC process took place from 8th April 2019 to the 5th June 2019, composed by several activities and events organised by the members of the ARCHIVER Buyers Group. As part of the process, a state-of-the-art analysis took place in the domain of archiving and digital preservation.

Archiving and digital preservation concerns the management of digital content over time to ensure ongoing access. It can be defined as "*the series of managed activities necessary to ensure continued access to digital materials for as long as necessary, beyond the limits of media failure or technological and organisational change*"[7]. This definition emphasises both the technical and organisational challenges involved in archiving and maintaining digital materials over time. In order for these materials to remain accessible in the long-term, they need to be managed as early as possible in their life cycle, preferably at the design stage. Targeted efforts in the research community are in place in order to determine the best practices which will ensure the sustainability of long-term data preservation.

As research and use of archives increasingly become digital, digital preservation is strategic interest for the public and private research sectors.

The following section provides details on this analysis, current and past initiatives and overall status.

[7] Definition adapted and updated from Beagrie, N. and Jones, M. 2001, Preservation Management of Digital Materials: A Handbook (British Library: London) p 10.

2. State-of-the-Art Analysis

The struggle of researchers to find affordable and sustainable archiving and preserving solutions for the medium to long term is compounded by multiple factors: funding models that are often limited to the lifetime of the research project itself, lack of clarity regarding available solutions, and no simple means to compare costs between different options.

In order to meet the goals of FAIR Data Management Plans (DMPs) for re-use and verification of results, preservation in this context encompasses the more challenging and advanced aspects of retaining processing and analysis software, documentation, and other components necessary for reusing a given dataset. Preservation can enable new analyses on older data, as well as a way to revisit the details of a result after publication. The latter can be especially important in resolving conflicts between published results, applying new theoretical assumptions, evaluating different theoretical models, or tuning new modelling techniques.

A number of research communities have been active in data preservation for some time, in some cases such as astronomy, even for decades. There are additional examples: the DCC[8] has developed a high-level overview model[9] of the stages required for successful curation and preservation of data from initial conceptualisation through the entire curation cycle; The Digital Preservation Coalition (DPC)[10] has raised awareness of the strategic aspects of data preservation and created several resources to help organisations to tailor best practices on data preservation to their specific needs. Other communities have come to the field more recently but still have significant experience and knowledge. Science Europe[11] for example has established a practical guide to allow the international alignment of research data management across Europe[12].

In spite of these longstanding efforts and accumulated expertise, in many scientific disciplines, findings are that most of the data of scientific activity is not yet published as it does not have the necessary metadata associated with it or data resulting from the majority of scientific activities cannot be found. In addition, interoperability across disciplines is usually not considered.

The European Open Science Cloud (EOSC) aims to address those issues for an estimated 1.7 million researchers in Europe. Services to make data Findable, Accessible, Interoperable and

[8] http://www.dcc.ac.uk/
[9] http://www.dcc.ac.uk/sites/default/files/documents/publications/DCCLifecycle.pdf
[10] https://www.dpconline.org/
[11] https://www.scienceeurope.org/
[12] https://www.scienceeurope.org/wp-content/uploads/2018/12/SE_RDM_Practical_Guide_Final.pdf

Reusable (FAIR), to store it and ensure long-term preservation are amongst the five main types of services to be made available to European researchers irrespective of discipline or national boundaries, forming the core of the EOSC design[13]. A set of aligned services for scientific data management, following best practices would form the base serving data FAIR to scientists, increasing data reuse and thus to a faster development of science.
This vision can lead to an effective European global structure: open, cost-efficient, as a result of standardization and best practices with trusted repositories supporting seamless data use across disciplines.
The ESFRIs[14] will also be engaged with the EOSC ecosystem. The ESFRIs role in the EOSC is very relevant as it can ensure stewardship of the data produced, encouraging scientific data sharing across their reference communities, to make it open as possible and widen the user-base to industry. The ESFRI landscape analysis[15] report states that:

> *"The multi-disciplinarily character of data-intensive themes imposes the scalability of the digital infrastructures with increasing e-needs, calling for a coordinated effort of all RIs. ESFRI foster adoption of FAIR – Findable, Accessible, Interoperable and Reusable – data principles plus data Reproducibility and Openness by all RIs of the Roadmap. ESFRI RIs generate massive data and have often developed own standards and metadata formats, developed data analysis and computational resources available to users, as well as data repositories for storage facilitating data sharing and re-use optimized for their reference scientific communities."*

The initial set of use cases present in ARCHIVER derive from four ESFRI landmarks[16]. Therefore, the resulting services of ARCHIVER, with R&D co-performed with ESFRI landmarks will help fulfill the archiving and data preservation requirements of most of the ESFRI projects and landmarks. This represents a significant market potential for the companies participating in ARCHIVER PCP.

The H2020 EOSC-hub project deliverable "D12.1 Procurement requirements and demand assessment"[17] highlights the importance of data preservation in the EOSC and recommends to

[13] https://www.eoscsecretariat.eu/sites/default/files/sip.pdf
[14] https://www.esfri.eu/
[15] http://roadmap2018.esfri.eu/landscape-analysis/section-3/big-data-and-e-infrastracture-needs/
[16] http://roadmap2018.esfri.eu/media/1049/roadmap18-part3.pdf
[17] https://documents.egi.eu/public/RetrieveFile?docid=3466&version=1&filename=EOSC-hub%20D12.1%20FINAL.pdf

collaborate closely with ARCHIVER on data archiving and preservation services. It also emphasizes the relevance of procurement instruments to democratize access to services (regardless of whether they are public or commercial) recommending that a procurement body, enforcing standards, as part of the EOSC entity, can work both as a de-risking factor and a quality assurance body for commercial services.
The H2020 project Big Policy Canvas[18] is developing a roadmap that will enable public administrations to improve their readiness with regard to the integration of Big Data. In its recent document, "Roadmap for Future Research Directions Research Challenges"[19] it highlights the provision of high quality, cost-effective, reliable preservation and access to data, as well as the protection of property rights, compliance with privacy and security standards for data as a research challenge for Open Government Data.

The cost of archiving and data preservation will have an impact on the long-term business model for EOSC. While considerable progress in the Member States regarding research data management (RDM) and open access to research data (ORD) is observed, the EC report on "Access to and preservation of scientific information in Europe"[20] highlights that

> "*National governments, as well as funders and institutions, are working already on various aspects of research data (data management, implementing FAIR, data management costs, etc.); varying levels of progress in relevant policies and practices can be observed across Europe. Hence, more coordination will be necessary to expedite process and align policies and practices*",

also stating that

> "*More work can be done with respect to funding of the costs of RDM . Nearly half of the countries report that national or institutional funding is available for RDM, while a third rely on EU funding for data management. Still, in many national funding schemes ORD related costs are considered eligible.*"

[18] https://www.bigpolicycanvas.eu/

[19] https://www.bigpolicycanvas.eu/sites/default/files/roadmap/BPC_Research_Challenges.pdf

[20] https://publications.europa.eu/en/publication-detail/-/publication/676f8a3b-62f6-11e8-ab9c-01aa75ed71a1/language-en

The European landscape emphasises both the importance of data archiving and preservation and that significant savings, high return of investment, optimisation and increased efficiency could be achieved if, on the one hand, commonalities and best practices across communities are exploited and, on the other, technical measures are implemented so conformance with current legislation is sought. As a consequence, this will lead to services that can be shared across communities (not necessarily via shared instances) and reduce fragmentation.
Considering the combination of these aspects, ARCHIVER resulting services will be very pertinent for research data preservation in Europe. ARCHIVER will develop and test services implementing best practices both for curation and reuse of the data resulting from scientific work, supporting effective research data management across different disciplines based on trustworthy repositories, where sustainability of the funding model is a critical factor. Cost models, in particular, are a subject of active discussion between the stakeholders of the project. ARCHIVER's results will be integrated into the future EOSC service catalogue with transparent "total cost of service" studies that can provide cost-effective solutions relevant at national, institutional and scientific community level.

2.1. Lessons Learned from similar activities

As part of its community landscape analysis, ARCHIVER is also identifying key lessons learned from other initiatives in the field. A particularly relevant project is The Digital Preservation Network (DPN). DPN was an R&D activity for on-premise digital preservation services, active from 2012 until 2019, initiated by a group of US university libraries and later set-up as a legal entity. DPN created a service to provide dark replication of data for institutions[21]. This system was designed to utilize established geographically separated repositories with emphasis on security, data protection, fixity checks and audits, as well as elimination of all single points of failure in the institutional infrastructures.
DPN has produced a final report[22], explaining the motivation for its creation, the activities carried out, the business model foreseen and why it failed, including the reasons for failure. Specifically, the report highlights that it is essential for all the procurers to agree on the services being developed and that each procurer must decide who in their organization is responsible for determining what information will be preserved. It also shows that the availability of lower priced, cloud-based solutions challenged the DPN services, and the short-comings of its business model. In a more technical perspective, bandwidth allocation to connect to the

[21] https://er.educause.edu/articles/2013/8/the-case-for-building-a-digital-preservation-network
[22] https://osf.io/3p9jq/

provided service has been overlooked and proved to be insufficient to handle large uploads of data within a reasonable amount of time. These reasons combined with the lack of user adoption on the services resulted in a significant decline in its memberships leading to the closure of the Digital Preservation Network in February 2019.
Some of the companies that were contracted to develop DPN on-premise services have participated in the ARCHIVER Open Market Consultation.
There is a number of lessons learned from DPN that the ARCHIVER project is taking into account by putting in place concrete actions to address them, in particular:

> "*Realistic and adaptable financial modeling was needed. The model placed too much reliance on potential deposits and payments.*"

ARCHIVER will require contractors across the three phases of the project to provide a Total Cost of Services (TCS) study of the future resulting services based on the architectural design proposed and its evolution during the project. The TCS aims to be a comprehensive study of all quantitative and qualitative factors significantly impacting the cost that a potential customer bears when purchasing and running the services resulting from ARCHIVER. The TCS will evolve with the project as the R&D is carried out: initial TCS at the design phase, updated TCS by the end of the prototype phase and final TCS by the end of the pilot phase. The objective is to establish a direct and consequent link between technical measures and architectural decisions in order to understand its influence in the final cost of resulting services.

> *"Product time-to-market was slow at the same time that member service needs changed rapidly"*

In order to reduce time-to-market, ARCHIVER is focused on using a staged approach for R&D. R&D is encapsulated across 4 layers (see Figure 1). The project aims are ambitious but realistic given the constraints of budget and time of execution.
Taking this into account, Layers 1-3 are considered the minimum R&D levels to be demonstrated. Layer 4 concerns the more advanced services such as data and software reproducibility and domain specific visual representation of data. As needs evolve and change rapidly, if R&D progresses and is quickly demonstrated in Layers 1-3, the ARCHIVER consortium will consider Layer-4 as "bonus R&D" for suppliers with a level of readiness allowing to provide those added value services.

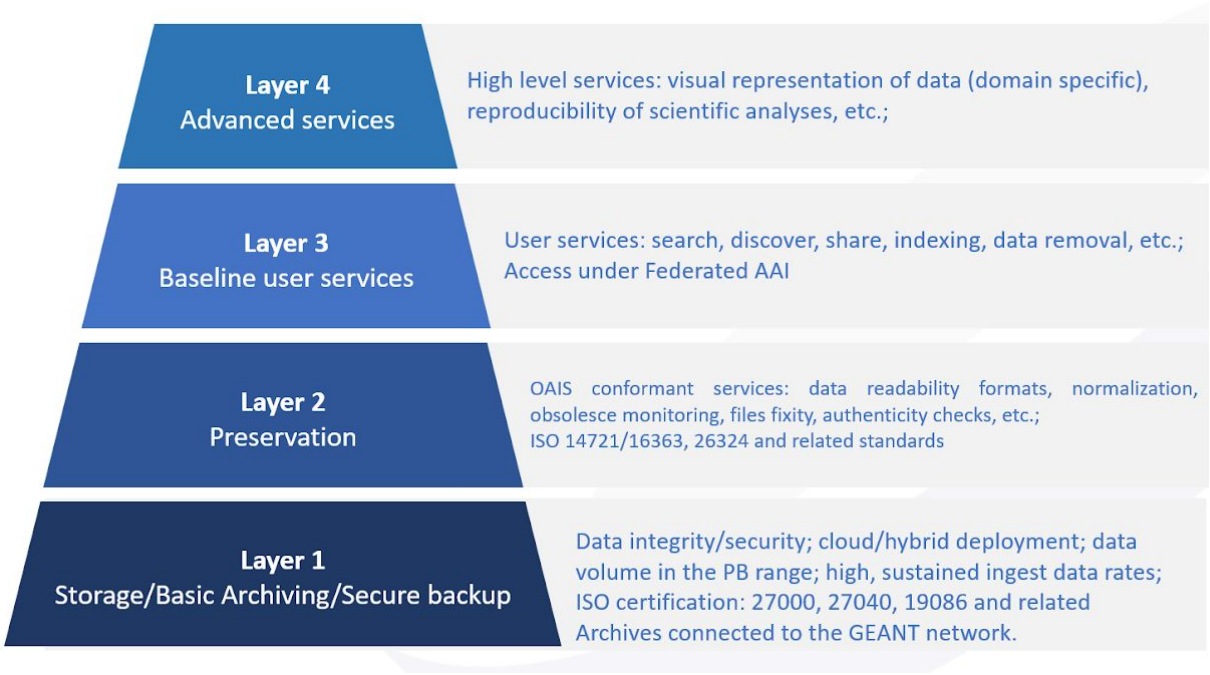


Figure 1: High level architecture for the future ARCHIVER archiving and preservation services.

> *"More DPN involvement with services for members and a focus on member engagement was needed. A focus on member governance should have been sought from the beginning."*

The Early Adopters programme (see section 3.1 of this document), was launched in order to encourage wider engagement outside the Buyers Group, to any public organisation having a need for innovative digital archiving and preservation solutions. The organisations taking part in the programme will be listed in the Request for Tender (RfT) document or in the subsequent call-offs and will have access to R&D that the contractors make publicly available. Early Adopters will also have a legal basis for procuring pilot-scale services directly from the ARCHIVER contractors and request the Buyers Group to obtain pricing for these services on their behalf under certain conditions.
The role of the Early Adopters compared to members of the Buyers Group is defined in the Request for Tender (RfT) document.

> *"Policies and service offerings were not agile and did not track quickly enough on changing needs"*

The integration of the services in the EOSC catalog will actively define technical, operational, legal and financial conditions for service providers to offer the future ARCHIVER resulting services. Specifically, the expectation is that the rules of participation in the EOSC will be published during 2020, covering aspects such as:

- Operational: defining service management requirements
- Technical: defining functional and interoperability requirements
- Legal: defining compliance requirements with current legislation (e.g. GDPR)
- Financial: establishing cost models and sustainability requirements

> *"Increased transparency in all directions among all partners including DPN Central staff, the Nodes, and Duraspace, would have improved communication and trust as the network developed."*

The objective of the Buyers Group in co-testing and co-designing a set of services for archiving and digital preservation during the ARCHIVER project is to reach a level of professional co-operation and trust with eventual contractors. Even if members of the Buyers Group and Early Adopter organisations will retain local copies of data, these organisations are entrusting service providers with valuable assets (data) that should always be considered irreplaceable. Transparency and rigor are necessary conditions on the R&D process, on costs of preservation associated services and respective overheads. The latter will always be higher when compared to basic data storage services for example. These conditions and overall approach will be reflected and taken into account throughout the project, notably during the assessment of each of the ARCHIVER R&D phases.

3. Open Market Consultation Roadmap

In preparation of the ARCHIVER PCP call for tender, a series of activities for both demand and supply sides have been held between April and June 2019.
The outcomes of this consultation are serving as a base for the tender specifications to be published in October 2019. The OMC was intended to be an evolutionary process, comprising parallel initiatives focusing on different user communities and stakeholders.
The OMC timeline is shown in Appendix A.

3.1. Community Requirement Landscaping

The ARCHIVER project team carried out various activities aiming at identifying the potential beneficiaries of the project at large and assessing their requirements in terms of digital archiving and preservation. An overview of the project stakeholders is available in Appendix B.

Collaboration with the Digital Preservation Coalition (DPC):

The Digital Preservation Coalition (DPC) is an international not-for-profit organisation, that aims to secure the digital legacy. It does so by enabling its members to deliver resilient long-term access to digital content and services, helping them to derive enduring value from digital assets and raising awareness of the strategic, cultural and technological challenges they face. Currently, 25 organisations, including CERN, are full members of the coalition while more than 70 joined as associated members.

The DPC agreed to collaborate with ARCHIVER, granting access to the Coalition's Knowledge Base as well dedicated support from DPC subject specialists in training and procurement of digital preservation services throughout the project lifetime.

During the Open Market Consultation, the DPC, in collaboration with the ARCHIVER project team and partner Trust-IT, organised a 1-hour tailor-made webinar training on Digital Preservation and Open Archival Information System (OAIS)[23]. This webinar was targeted to the end-users of the digital preservation services that will be developed during the project and to the community at large.

A total of 18 persons followed the live webinar, from which 13 are potential end-users and 5 are potential tenderers (see Appendix C for details). In addition, the recording of the webinar was published on the project website and watched 38 times. DPC involvement in ARCHIVER may evolve to the provision of end-user training on the resulting ARCHIVER services.

Participation in the EOSC-hub week:

The ARCHIVER project was presented at the EOSC-Hub week that took place on the 10-12 April 2019 in Prague, Czech Republic. It also participated in a panel discussion about the role of cloud services procurement in the EOSC context. The event raised awareness of the project generating a lot of interest. The audience considered the topic of archiving and preservation of importance, especially in the EOSC context where no other project is addressing it as a core activity. The audience specifically mentioned the problem of the current in-house archiving and

[23] https://www.archiver-project.eu/training-digital-preservation-and-open-archival-information-system-oais

preservation services, essentially based on tape technology with all the costing risks associated to it due to market monopoly, taking into account the global scenario[24]. Interest has also been expressed by potential ARCHIVER future adopters including National Research Networks (NRENs), International Research Organisations and National Libraries.

Participation in iPRES2019:

iPRES[25] is the premier and longest-running conference series on digital preservation. Since 2004, annual iPRES conferences are in rotation around the globe on four continents. iPRES brings together scientists, students, researchers, archivists, librarians, providers, and other experts to share recent developments and innovative projects. The discussions promoted at iPRES have evolved significantly over the last few years, from a digital preservation technology driven niche of experts to a community global challenge, covering a wide variety of topics in digital preservation from strategy to implementation, and from international and local initiatives.

ARCHIVER has been selected to present a poster on its activity during the project preparation phase, at iPRES2019 in Amsterdam 16th-19th of September.

Participation in IDCC2019:

Similarly to iPRES, ARCHIVER submitted a contribution to iDCC international conference[26] to be hosted in Dublin between the 17th-20th February 2020. iDCC, organised by the Digital Curation Centre (DCC), focus on the community and stakeholders that play a role in ensuring digital objects are properly created, managed and shared. The ARCHIVER contribution, if accepted, will report on the results of the tender evaluation for the Design Phase.

Involvement of GÉANT for network connectivity and Federated Identity Management (FIM):

In the context of EOSC landscaping activities, part of the EOSC ambition is to join existing and emerging horizontal and thematic data infrastructures, bridging fragmentation and leveraging on past infrastructure investment. Taking into account this scenario, the GÉANT organisation provides the high-bandwidth pan-European research and education backbone that interconnects National Research and Education Networks (NRENs) across Europe. It also provides worldwide connectivity through links with other regional networks.

24 https://www.theregister.co.uk/2017/06/14/spectralogic_foresees_ibm_becoming_the_sole_tape_drive_supplier/

25 https://ipres2019.org/

26 http://www.dcc.ac.uk/events/idcc20

GÉANT's role is very much linked to data movement and access in the EOSC: GÉANT and the NRENs provide not only European network connectivity but also a federated access and authentication service[27].

During the OMC, potential bidders have been exposed to these core requirements. In one of the OMC events, GÉANT representatives invited by the ARCHIVER project have exchanged with the supply side, in order to estimate the effort required to ensure large end to end network capacity (in multiples of 10 GB/s) between the suppliers' facilities and Buyers Group data centres using the GÉANT network. Information has been provided about GÉANT procedures and processes highlighting that GÉANT is not a commercial network operator.

A similar approach has been used to cover Authentication and Authorisation Infrastructures (AAIs), a core requirement in the EOSC context.

A considerable number of research and education organisations use AAI to build a trusted environment where users can be identified electronically using a single identity. This requires an Identity Provider (IdP) to be able to authenticate users and provide a limited set of attributes that characterise a given user in a given context.

During the OMC, potential bidders have been exposed to the AARC blueprint architecture[28] including the proxy services[29] available as production or pre-production services across the research community. The objective was to make potential bidders understand and estimate the effort needed to support inter-federation schemes in the context of the ARCHIVER resulting services. This dialogue is continuing with GÉANT, in order to offer a single interface to researchers, allowing them to connect Identity Providers (IdPs) from the national academic federations towards the future ARCHIVER services.

Launch of the Early Adopters Programme:

The ARCHIVER Early Adopters Programme was launched in order to encourage wide deployment of the ARCHIVER services outside the Buyers Group to any public organisation having a need for digital archiving and preservation solutions.

ARCHIVER aims to ensure that the resulting solutions are as widely applicable as possible.

Organisations that become part of the Early Adopter program will:

- Be consulted during the preparation of future ARCHIVER phases;
- Access material produced by the project;

[27] https://edugain.org/
[28] https://aarc-project.eu/architecture/
[29] https://wiki.geant.org/display/AARC/IdP-SP-Proxy+policy+framework

- Propose their own use cases that they are willing to deploy, test resulting services and assess the support for OAIS, FAIR and Open Data as well the cost-effectiveness;
- Benefit from training sessions covering the services developed during the ARCHIVER project.

In addition, ARCHIVER will be able to request firm pricing for limited pilot-scale use of any of the resulting services on Early Adopters' behalf, during the call-off for the pilot phase.

The legal basis for ARCHIVER Early Adopters being able to purchase limited pilot-scale services if they are contracting authorities subject to EU procurement rules is article 32(3)(a) of Directive 2014/24/EU[30]. If Early Adopters are not contracting authorities then they will have to check their own rules which will govern their purchase of these services.

The organisations taking part in the programme will be listed in the Request for Tender document or in the subsequent call-offs.
At the date of publication of this document, four organisations formally expressed interest in the programme: Jisc (National Research and Education Network in the United Kingdom)[31], joined to address its member's big data, high throughput and large file size preservation issues. SURFsara (National Research and Education Network in the Netherlands)[32] and SWITCH (The Swiss National Research Network)[33] both aim to provide solutions for long-term data preservation and also for disaster recovery and business continuity off-site of their campuses. In addition, the New York State Canal Corporation[34], a New York state government agency (US) that runs the New York canal system, joined as an Early Adopter. NYSCC collects data to analyze and predict water levels in the NY canal system. The predictions allow to determine commercial, agricultural and traffic plans including forthcoming operations. Data is used by the public and government to inform executive decisions on state investments. The objective of NYSCC is to replace their current data share drive as its archive, that incurs risks of redundant, outdated, and trivial data that wastes time and resources, slowing down their data science efforts. NYSCC is looking for an effective, efficient, and compliant data archiving tool that would help speed the time to action by lowering the time to find and increasing the data trust level.

[30] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32014L0024
[31] https://www.jisc.ac.uk/
[32] https://userinfo.surfsara.nl/
[33] https://www.switch.ch/
[34] http://www.canals.ny.gov/

There are several other organisations considering the Early Adopter programme: these include National Libraries, Universities and International Research Organisations. National Repositories have been contacted by the project team via the OpenAIRE project including data curation expert communities such as the DCC.
To continue to raise the broader community awareness about the programme, the project team will organise a first information webinar in September 2019[35].

3.2. Dialogue across Demand and Supply sides

Preparation workshops:

In order to assess the innovation potential of use cases in a pre-commercial procurement, two dimensions need to be considered: the added value for the end-user, and the estimated risk for the supplier. ARCHIVER performed this analysis using the "planning poker" technique. Planning poker is a "best practice" methodology to estimate added value, level of complexity, required implementation effort or risk. The technique is based on domain expert evaluation and achieving consensus. It uses a Fibonacci sequence to reflect the inherent uncertainty in estimating larger items. The application of the technique was moderated by the consortium member Addestino.
Procurers were asked to estimate the value of a specific use-case using the scale represented in Figure 2.

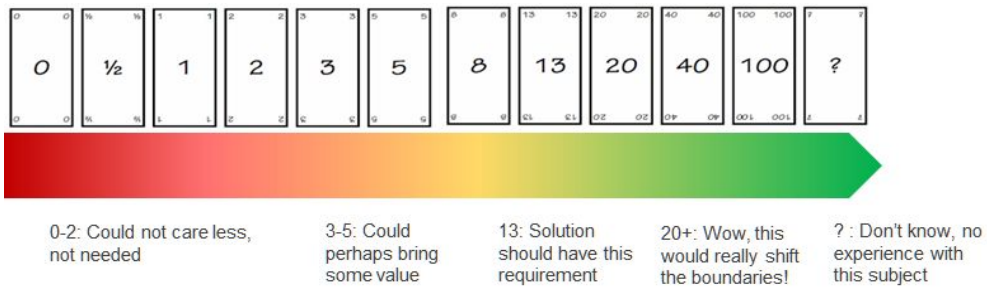


Figure 2: Added Value Assessment Scale.

Suppliers were asked to estimate the risk of a specific use-case using the scale scale represented in Figure 3.

35 https://www.archiver-project.eu/early-adopters-programme-webinar-26-june-1500-cest-new-date-4-september-2019

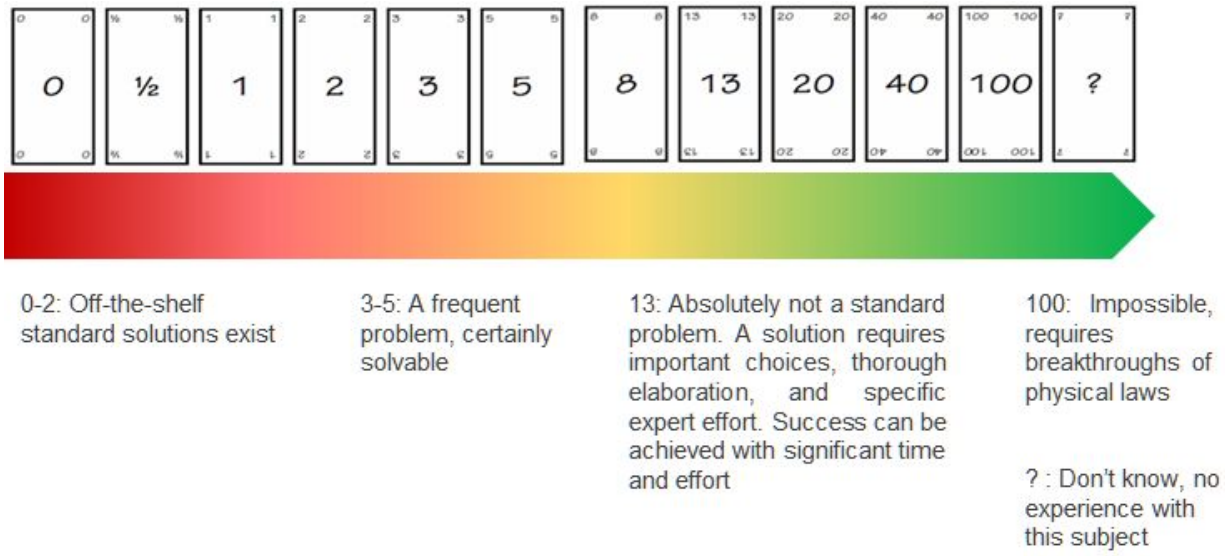


Figure 3: Risk Assessment Scale.

By combining both scores, it is possible to assess the innovation potential of a use case. This is done by plotting each use-case on a two-dimensional graph. The vertical axis represents the added value for the end-user whereas the horizontal axis reflects risk.

Within the PCP, the highest priorities for the envisaged solution are the items that realize the highest added value for the end-user, while holding a moderate to high risk.

In order to implement this technique, the procurers made a first analysis scoping of the different requirements of the project. From these requirements, different atomic use-cases deriving from the Buyers Group deployment scenarios were identified and prioritized based on their value for the end-user based on the scale depicted in Figure 2. Each use-case follows a distinct pattern ("*as an <actor> I can <capability>, so that <reason>*") identifying the targeted audience, the aspired capability and the reason for its need.

This classification was done during two workshops: on the 8th of February 2019, immediately after the project Kick-off meeting at CERN in Geneva and on the 20th of February 2019 organised by DESY in Hamburg.

The outcome of the two use-case scoping sessions was a list of 47 atomic use-cases prioritized based on their added value for the end-users. The atomic use-cases were listed in 7 categories:

- OAIS Conformance
- Hybrid Archives
- Business & Deployment Models
- Scalability

- Federation & Integration (of archives)
- Network & Federated AAI

This list of the atomic use-cases is available in Appendix D.

Open Market Consultation events:

A series of supplier workshops actively approaching the market to obtain information about the current state of the art, developments and innovation potential in the archiving and preservation sector has taken place during the ARCHIVER OMC.

The dialog started on the 8th April 2019 with a series of four events targeted to potential tenderers and open to end-users. In order to facilitate the participation of companies and end-users and reduce the environmental impact linked to travelling, the events were organised by the Buyers Group in three different European countries: Spain, United Kingdom and Switzerland. These events were promoted by TRUST-IT both on the project website and social media channels. In addition, the Buyers liaised with different local organisations with the aim of increasing the events' visibility. PIC cooperated with the Catalonia Trade & Investment Agency (ACCIÓ) and Europe Entreprise Network (EEN) for the venue and the dissemination of the event. Similarly, the Stansted event was promoted by two EEN Contact Points in London: Newable Limited[36] and the University College London[37]. CERN liaised with its Industrial Liaison Officers (ILOs)[38] to communicate about the supply-side events and made a publication on its procurement website[39].

These dissemination actions proved to be very efficient: more than 170 people participated in the events and 38 different companies from 14 countries were represented. The number of participants at the OMC events and the geographic breakdown of the organisations attending are available in Appendix E and F.

The majority of these companies are SMEs specialized in the following fields of activity (listed per order of importance):

- Data Preservation and Archiving Services
- Cloud Services Providers
- Consulting Services

[36] https://newable.co.uk/
[37] https://www.ucl.ac.uk/
[38] https://procurement.web.cern.ch/en/who-to-contact-in-your-country
[39] https://found.cern.ch/java-ext/found/CFTSearch.do

Details on the size and the type of companies that participated in the events are available in Appendix G and H.

In addition, the events were attended by many end users from the Buyers Group and by several representatives from public organisations in need of innovative data preservation services. The list of companies and organisations that followed the events is available in Appendix I.

As the process was designed to be evolutionary, at each event, the participants were given the opportunity to provide their feedback and the focus of the following events was adapted based on the feedback received. The summary of the feedback received is available in Appendix J.

The first two events were organised by CERN in Geneva on the 8th of April and by PIC in Barcelona on the 7th of May. Both workshops were moderated by Addestino and using the planning poker technique, the industry representatives were asked to rate the risk associated to the atomic use-cases identified by the Buyers Group in the preparation workshops. The outcome of these two workshops has been a detailed analysis of the value versus the risk associated to each atomic use case. This analysis is summarized in section 3 of this report.

The planning poker was a valuable exercise in order to scope the debate around the R&D effort needed for the project with a significant level of detail. At the same time, the feedback received after the second workshop showed that the industry was lacking an overview of the challenge to be addressed and additional technical information on the possible future market opportunities and applications. Therefore, the project team decided to adapt the format of the forthcoming OMC events in order to address the supply-side need for information.

The following supply-side event was organised by EMBL-EBI in Stansted, United Kingdom on the 23rd of May. In order to provide the supply-side with a more concrete understanding of examples of the future market needs and applications of the R&D solutions to be developed in the PCP, the Buyers Group provided detailed technical information about their deployment scenarios.
In addition, ARCHIVER invited experts from GÉANT, in the areas of networking services and Federated Identity Management (FIM) in order to describe these services and possible options in the context of the European research and education community and in the specific

implications for the project. The fruitful discussion during the workshop helped the Buyers Group to identify the aspects of the project that are considered challenging by the industry.
The industry's feedback after this event showed that potential tenderers experienced some difficulty in understanding the common characteristics of the various deployment scenarios. There was a certain degree of divergence perceived across the different deployments and concerns were expressed if one single solution would address all the needs.
Potential tenderers asked for clarification on the weighting criteria in order to understand which requirements are prioritized by the Buyers Group.

The project team digested the feedback and clarified this aspect at the OMC Consolidation event that took place on the 5th June at CERN. Specifically, a mapping of draft versions of the selection and award criteria into the deployment scenarios has been presented and discussed.

Overall, the series of events allowed to fine-tune the R&D potential as tenderers and end-users came together and discussed the deployment scenarios requirements. This open dialogue resulted in an analysis of the R&D scope of the project which is summarized in section 3 of this document.

Frequently Asked Questions (FAQ):

In order to ensure all stakeholders benefit from an equal access to the information related to the project and the tender exercise in preparation, all the answers given to questions raised during the OMC process were documented and published on the project website in the form of an FAQ[40] for public access. This webpage was regularly updated as new questions were received. In addition, organisations were given the opportunity to ask additional questions via a form available on the webpage.

Draft PCP Contract Notice Feedback:

In order to stimulate an open and transparent process, the draft PCP Contract Notice has been made publicly available on the project website[41] for consultancy. Potential tenderers were invited to consult all documents and encouraged to comment on the Functional Specifications, the Qualification and Evaluation Questionnaires, the Financial offer and the list of Deliverables and Milestones for contractors during the implementation phases.

The project team was particularly interested in feedback on the proposed scope of the R&D challenge and the feasibility and risk for companies to participate in the project.

[40] https://www.archiver-project.eu/faq
[41] https://www.archiver-project.eu/draft-tender-documents

A first version of the draft PCP Contract Notice was published on the 3rd May. An updated version of the draft document was published on the 6th June, following the OMC Consolidation event. This updated version already addressed some of the feedback received on the documents, including:

- comments related to the tender process in particular concerning the restrictions in place regarding having multiple bids by a lead contractor
- addition of an award criteria requesting more details about cybersecurity requirements
- ability to propose networking solutions in alternative to GÉANT

Based on the feedback received at the OMC events, potential tenderers representatives appreciated the transparency of the process. When asked if they have consulted the documents, all company's representatives stated that they either already had or planned to do so (see Figure 4). In addition, the project team has offered the participating companies to showcase their state of the art and current solutions publicly in the project website[42] or internally with the project team.

[42] https://www.archiver-project.eu/existing-archiving-and-digital-preservation-solutions

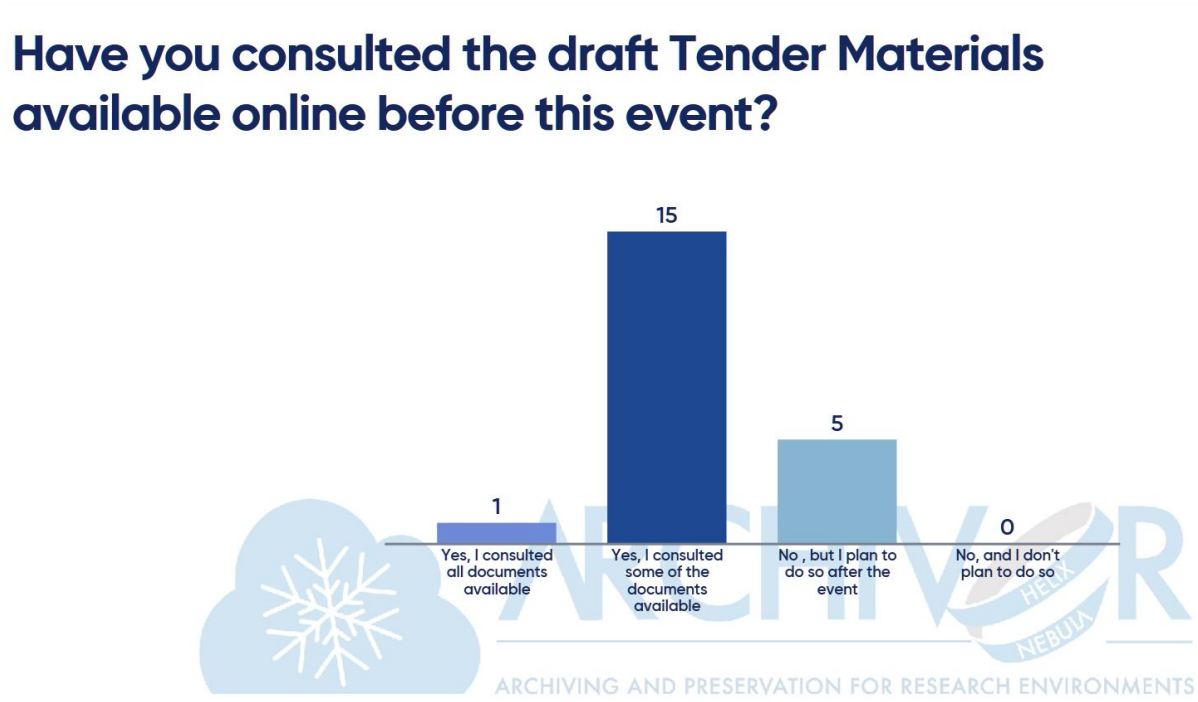


Figure 4: Extract of the feedback session of the OMC event in Stansted.

Consortia Matchmaking:

Since different skill sets are needed to cover the different challenges of the project, the project team has encouraged the formation of consortia and subcontracts in order to bid for the tender (Figure 5). In order to avoid any conflict of interest, the project team did not take an active role in this process but rather offered a voluntary means for the companies to look for partners.

After an analysis of the different tools and services available, the project team selected the online platform Crowdhelix[43] to serve this purpose. Crowdhelix is an Open Innovation platform connecting an international network of universities, research organisations & companies. About 2.5K collaborators, including 70+ universities from 40+ different countries and 300+ industry partners, are using the Crowdhelix platform.

[43] https://www.crowdhelix.com/

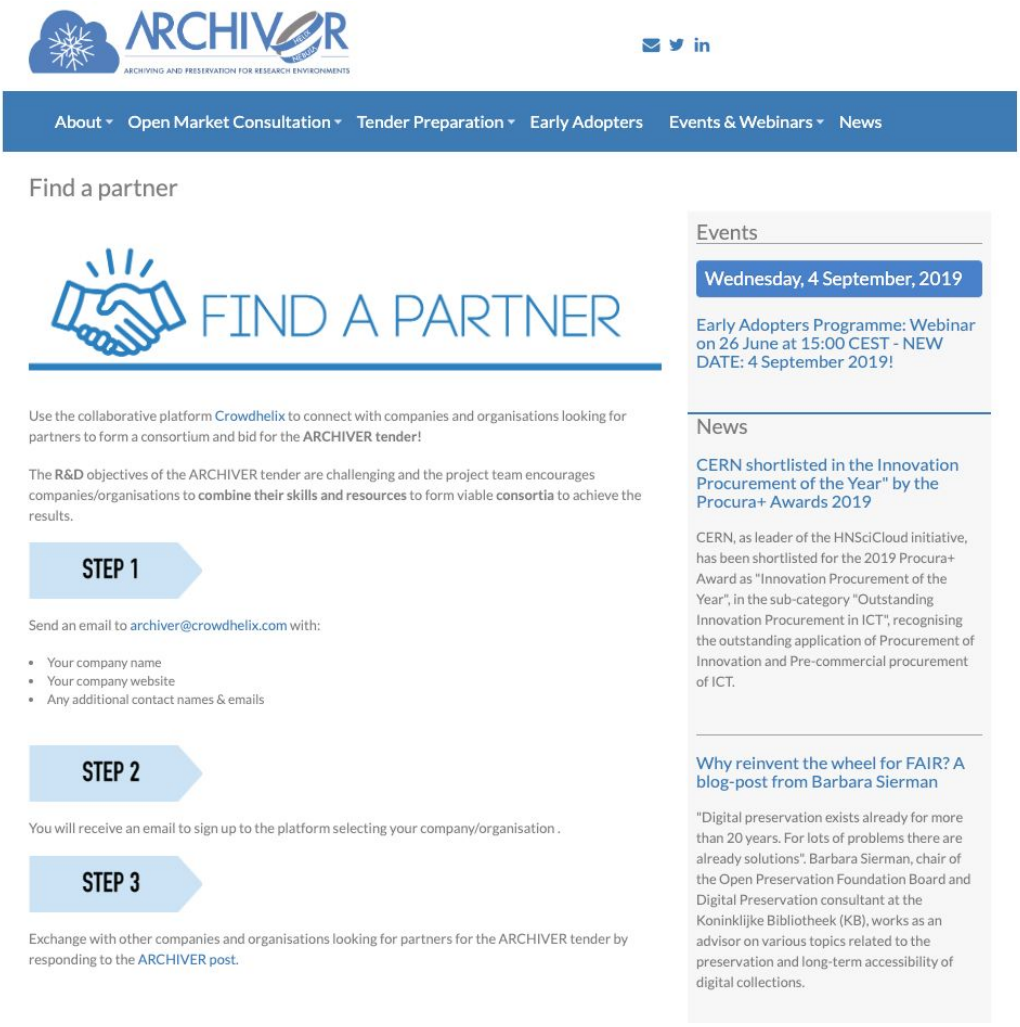


Figure 5: "Find a Partner" pager in ARCHIVER website using the Crowdhelix.

The platform works on a subscription based model in order to ensure its sustainability. For the purpose of the project, the Crowdhelix team kindly offered to make the platform available at no cost to interested companies.

In order to present and set up the platform, Crowdhelix held two video meetings with the ARCHIVER project office. In addition, the Crowdhelix CEO attended in person one OMC event to get additional knowledge about the project.

The platform has been promoted to the industry both at the OMC events and via the project website[44]. 14 companies signed-up on the platform and 7 public posts from companies looking for partners were published.

[44] https://www.archiver-project.eu/find-partner

The ARCHIVER consortium is not in a position to see the direct outcome and any activity beyond the public posts. The Crowdhelix team reported that it is usual that companies take the discussion off the platform once the first contact is established.
During the OMC Consolidation event, the ARCHIVER team received oral feedback from companies that the tool was useful but that an active moderator role would be appreciated. The project team discussed this possibility with the Crowdhelix team who reported being in a position of performing the role of moderator should the effort associated with it be funded.

For future projects and especially in the context of the EOSC, the use of such a platform with a moderator role would be greatly beneficial. The Crowdhelix team is pursuing this direction and is creating an Open Science Helix that will be officially launched in October 2019[45].

4. Main Outcomes of Open Market Consultation

4.1. Demand Side

Throughout the series of landscaping activities performed during the OMC, the community has shown growing enthusiasm for the project. A diversity of public organisations including National Research and Education Networks (NRENs), National Repositories, Universities, National Libraries, International Organisations, etc., acknowledged facing similar challenges when it comes to preserving their data. Many of these organisations are looking for disaster recovery solutions in order to ensure business continuity in case of shutdown of the main long term preservation archive system, also due to the risk associated with the current evolution of the main tape technology in use (LTO). In addition to the technical challenges, sustainability and cost-effectiveness of solutions developed during the project is of great importance, together with exit strategies and conformity with relevant European legislation.
In the context of the EOSC core objectives, ARCHIVER will have a key role of providing future services integrated and tailored to address the implementation of best practices for archiving, curation and reuse of the data resulting from scientific activities.

The R&D challenge of ARCHIVER has been encapsulated into four interdependent layers (Figure 1). This approach allowed to stage R&D in order to set expectations correctly with objectives that are measurable and realistic: while most deployments will require that the solutions developed provide functionalities in layers 1-3 as a minimum, layer 4 functionalities represent a bonus R&D layer, bringing increased added-value in many of the deployment scenarios.

[45] https://network.crowdhelix.com/events/2019/6/5/crowdhelix-rto-and-corporate-members-event

The mapping of a set of initial deployment scenarios onto the layered classification resulting from the OMC and relate community landscaping activities is depicted in Figure 7.

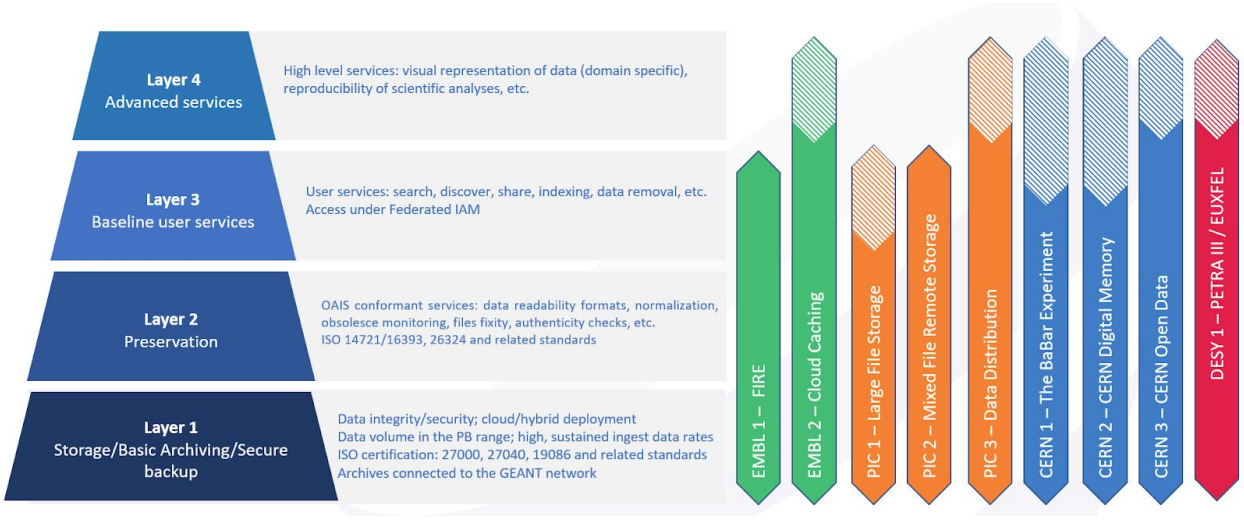


Figure 7: Mapping of the Buyers Group's deployment scenarios on to the high level architecture.

Solutions that map on to this high level architecture must consider two additional core aspects:

- Cost-effectiveness of the services
  A key requirement in order to allow the broader commercialisation of the solutions developed across the research community. The business model needs to take into account aspects such as the scale of data and ingests, archives lifetime and number of copies. As mentioned earlier, deployment models are of great importance, especially when they concern strategic components such as exit plans, in order to avoid vendor lock-in, a fundamental barrier in the adoption of cloud hosted preservation services.

- Adherence to European legislation such as the General Data Protection Regulation (GDPR)
  Solutions must demonstrate that adequate capabilities are supported and/or will be implemented as appropriate technical measures during the R&D activity, so that the resulting archiving and preservation services can handle correctly personal data present in scientific data sets.

4.2. Supply Side

The different initiatives targeted to potential tenderers during the OMC process helped to identify the R&D potential and evaluate effort and risk associated to the R&D to be performed.

Outcomes of the Planning Poker:

The planning poker technique helped the analysis of the value versus the risk associated to each atomic use case. The outcome is shown in Figure 8.

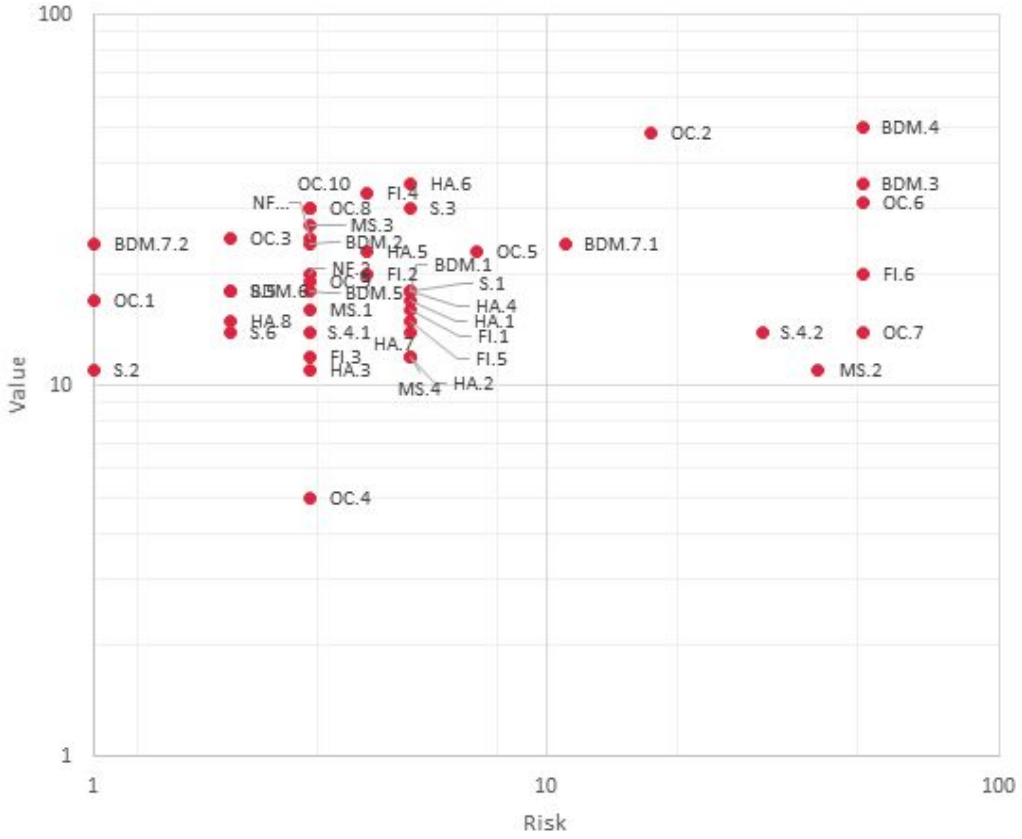


Figure 8: Atomic Use Cases Value versus Risk Matrix.

On the vertical axis, the innovation potential is displayed as assessed by the buyers group members. This is based on how valuable the atomic use-case would be for the different user groups.

On the horizontal axis the technological risk is displayed as assessed by the group of suppliers participated in the open market consultation.

In this matrix four distinct zones can be distinguished:

- Top left quadrant:

The critical functionalities can be found with high added value to the end users but low technological risk. The more the atomic use-case is positioned on the left, the more the market indicates an off-the-shelf solution exists or a similar use-case has been realised with well-established technology.

- Top right quadrant:

The functionalities with high added value to the users but high technological risk can be found. In this case the market indicates a solution needs to be found together with the Buyers Group in order to satisfy the specific needs. This solution will likely require new development or a proof of concept. The more a use-case is positioned on the right, the more likely there is a need for technological research.

- Bottom left quadrant:

Functionalities placed in this quadrant have a lower added value but also a low technological risk.

- Bottom right quadrant:

Functionality with a low added value imposing high technological risk. These use-cases are irrelevant for the final solution.

Summarizing, the requirements embedded in the atomic use-cases with high added value to the end-users and relevant technological risk for the industry (i.e. located in the top right corner of the matrix) can be grouped as follows:

- FAIR principles
- Scientific data management
- OAIS reference model
- Portability of archives and exit strategies
- Replication of archives on multiple sites
- Reproducibility of the scientific analysis
- Data volume scalability and very high sustained ingest rates
- Management of personal data

Outcomes from the OMC events:

The exchanges that took place during the OMC events were crucial to provide information about the research community requirements and to distinguish functionality already available in the market from that requiring R&D effort from the industry. Technical summaries[46] for the initial list of deployments of ARCHIVER were published in order to maximise the amount of structured information available to potential tenderers.

Figure 9 shows the feedback received from potential tenderers when asked about the R&D effort inherent to the deployment scenarios foreseen in the project.

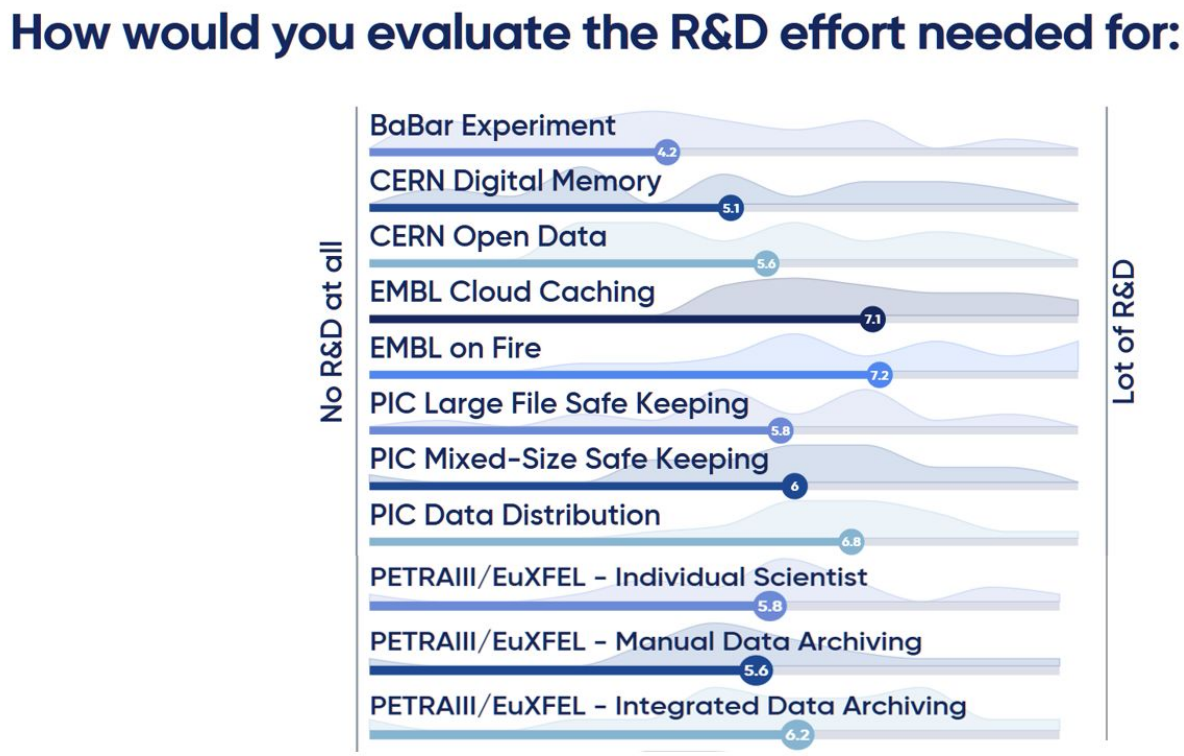


Figure 9: R&D evaluation by the companies.

While industry identified R&D effort in every deployment, a higher level of R&D is needed for:

- EMBL on FIRE[47] (estimated R&D effort: 7.2/10)

[46] https://www.archiver-project.eu/deployment-scenarios-technical-summaries
[47] https://drive.google.com/file/d/12-ZXK_2_OiNZz-l0HZ7QBw9VLMREnVFJ/view

- EMBL Cloud Caching[48] (estimated R&D effort: 7.1/10)
- PIC Data Distribution[49] (estimated R&D effort: 6.8/10)

These deployment scenarios have some exceptional requirements:

- Scale: as an example, the archive of EMBL is 20 PB in size, and is currently doubling every two years. This growth rate is expected to continue for at least the next decade.
- Frequent and complex access patterns: the data needs to be frequently accessed by users scattered across multiple geographical locations. This requirement directly influences the network bandwidth necessary during the project.
- Data privacy: some of the data used for the EMBL Deployment Scenarios contains important Personal Identifiable Information (PII), being a recognised R&D challenge to have technical measures in place to ensure personal data is processed under applicable legislation.

In addition, industry estimated that the effort needed to meet the Network Connectivity and Federated Identity Management requirements for the deployment scenarios is significant (see Figure 10).

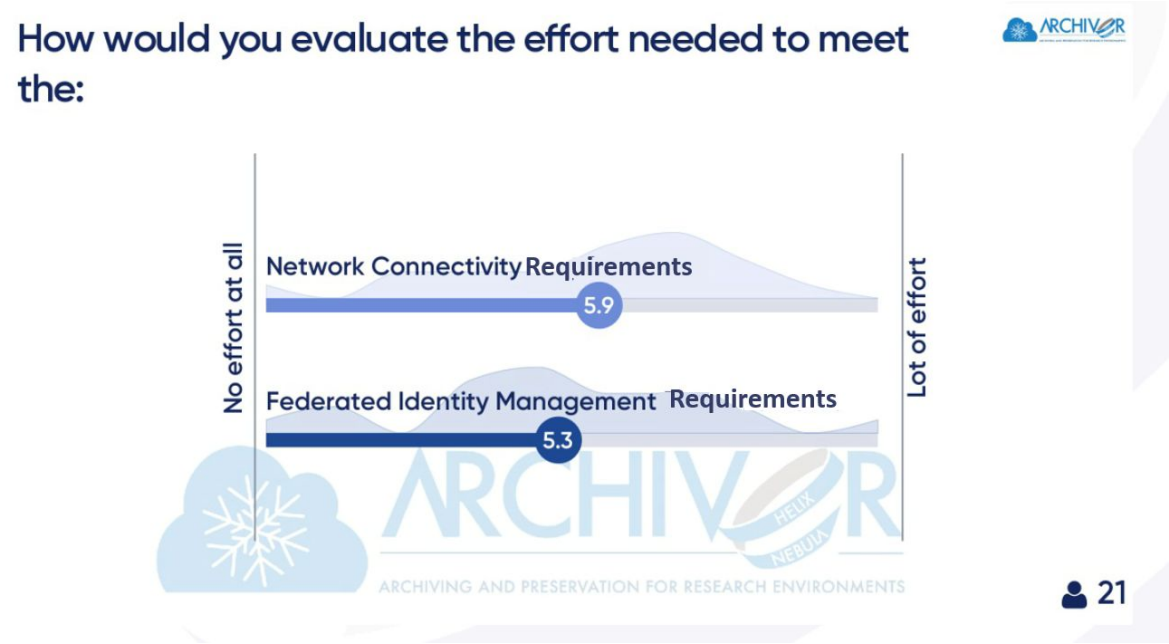


Figure 10: Effort estimation for GEANT network connectivity and FIM requirements.

[48] https://drive.google.com/file/d/120VjL44fD7xYziTf5UAPGLIjcETy8dwF/view
[49] https://drive.google.com/file/d/11i6aN38T1_ME43H9_4Jk7JIXy_o7B-tO/view

Finally, when asked to rank functionalities based on the R&D effort associated, the industry ranked FAIR principles, OAIS Reference Models, Data Privacy and the associated Business Models first (see Figure 11).

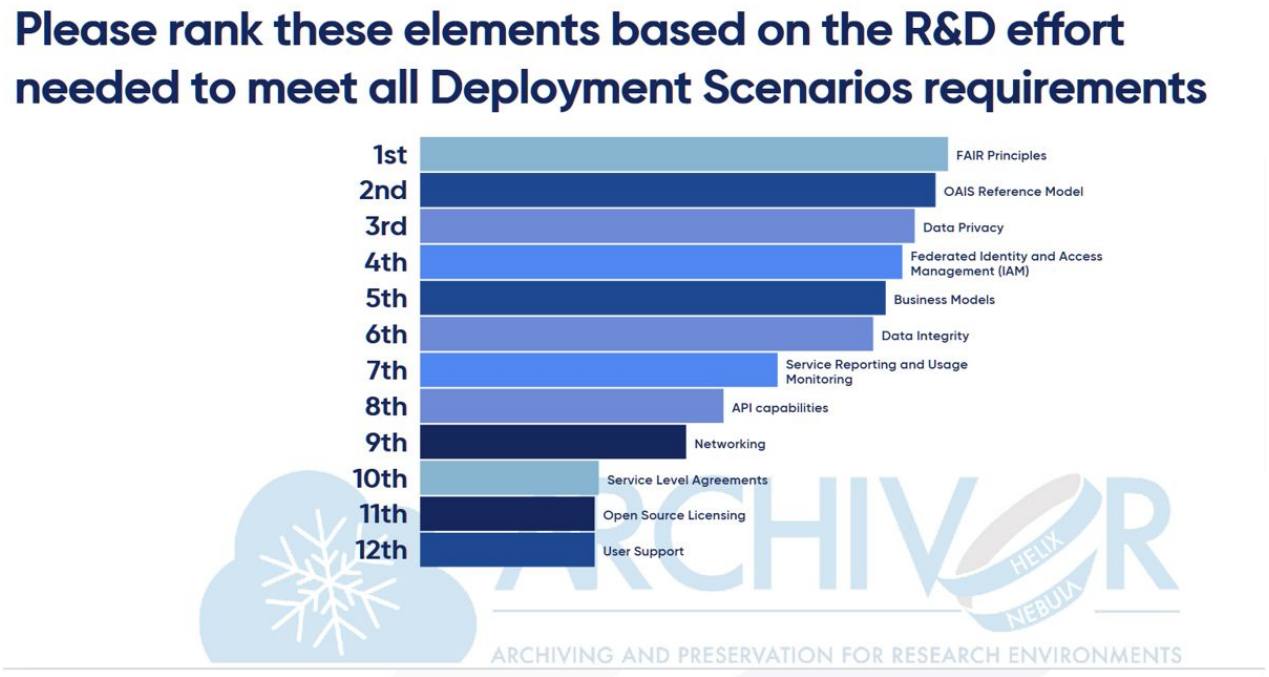


Figure 11: Ranking estimation for R&D effort per topic.

The outcome obtained with the planning poker method and the discussions during the OMC events show that the following requirements cannot be met without R&D effort from the industry:

- Business Model of the solution
- Data privacy
- FAIR Principles
- Federated AAI
- Management of scientific datasets
- Network Connectivity
- OAIS Reference Model
- Scalability

5. Technological baseline state of play

5.1. Cloud Market

During the preparation phase, the ARCHIVER consortium has analysed what solutions are available in the market. In general terms, commercial services in the area of archiving and data preservation are currently available however, they have not yet demonstrated to scale to the petabyte region and beyond, nor address the scientific data types, often domain-specific, that are needed by many scientific disciplines.

All providers of infrastructure services – both commercial and public – are able to offer cloud-based storage solutions, but there are limitations when it concerns archives. Archives typically have additional requirements beyond the simple availability of a place to store data files. These requirements may include specific concerns around data protection and the processing of personal data, or even mitigation measures for data loss.

According to Radicati[50], the total worldwide Information Archiving market will be over $5.5 billion in revenues by year-end 2019 and will grow to over $8.8 billion by 2023. In the table below is a list of some of the most well-known commercial suppliers and their products in the market today:

Amazon AWS Glacier	Ex Libris	OpenText InfoArchive
Artefactual Archivematica	Google Vault	OVH Public Cloud Archive
Arkivum PERPETUA	Iron Mountain	Preservica Cloud
Atempo Digital Archive	Jatheon Archiving Suite	Proofpoint Enterprise Archive
Autonomy	Micro Focus International Digital Safe	SMARSH Archiving Platform
Barracuda Cloud Archiving	Microsoft Azure	Veritas Enterprise Vault
DuraCloud	Mimecast	

Table 1: Leading suppliers in the cloud-based information archiving market.

Cloud-based archiving solutions provide a number of features such as deployment options including combination of several types such as private, public, hybrid or even community and special purpose clouds;[51] with support for multiple content sources and definition of data

[50] https://www.radicati.com/?p=15991

[51] https://ec.europa.eu/digital-single-market/en/cloud-computing-expert-group-research

retention policies. However, beyond these features, a number of considerations for cloud-based solutions need to be taken into account:

- Cloud archiving is a trade-off between accessibility and cost. Lower cost means limited data access performance.

- Using a cloud provider requires the research organisation to shift its internal workflows: from managing machines to managing service levels.

- Research organisations need clearly defined service levels for cloud-based archive hosting to be able to ensure a high-quality and sustained end-to-end service for their users.

- Research organisations must have an exit strategy in case there is an issue with the cloud-based service provider.

- Security and integrity of data is essential for any data storage solution. Hyperscale cloud service providers invest significant sums in ensuring the security of their services and are usually certified according to standards such as ISO 270018 and 270029 that describe the steps to be taken in maintaining physical and online security, and how to respond to breaches.

- External cloud services are considered to be operational rather than capital expenditure: instead of buying computer equipment up-front and writing its cost down over several years, customers of cloud services are billed for the computing power, network bandwidth and storage space that they actually consume. This can be cheaper for research organisations, especially for short or fluctuating workloads, but it requires them to think differently about the way their budgets are planned and managed.

In addition, specifically for the public research sector, three elements are of great relevance:

1) Data held in archives must be expected to be both preserved and accessible beyond the commercial lifespan of any current technology or service provider;

2) An approach to addressing serious risks, such as loss, destruction or corruption of data that is based purely on financial compensation will not be acceptable, as this takes no meaningful account of the preservation and custodial role of archives;

3) In order to reinforce the criticality of the first two elements, explicit provision must be made for pre-defined exit strategies (e.g. agreeing in synchronising data that can be restored across two different service providers, or between a cloud provider and

on-premise storage; or setting up an escrow service[52]) with effective monitoring and audit procedures.

In terms of R&D activities not yet close to market, some hyperscale commercial cloud providers, are performing innovative activities on cold storage technology[53] with the aim of replacing both tape and optical archival discs as the media of choice for large-scale and (very) long duration cold storage in order to reduce costs and increasing reliability in its own cold storage programs.

5.2. Existing commercial data preservation and archiving solutions

The Digital Preservation Coalition (DPC) provides a body of material[54] presenting the current state of popular preservation and archiving products available on the market, as depicted in the table below.

Digital Preservation Futures Webinar	Provider
Episode #1	Arkivum
Episode #2	MirrorWeb
Episode #3	LIBNOVA
Episode #4	Formpipe
Episode #6	Preservica
Episode #7	Artefactual

Table 2: Archiving & Preservation specialist companies webinars.

The ARCHIVER consortium has studied this material to assess the current state of the art.
The material highlights the existence of modular systems which may be adapted to the customers' needs. This approach requires less development effort but more configuration, integration and adaptation processes in order to fully address the use cases. The observed solutions serve specific file types only, however there are examples which purport to process multiple input formats. Furthermore, the compliance level of the solutions varies and needs to be adapted to fit the public research use cases and European regulation. In what concerns scale, the ingestion rates and service availability have never been tested up to the tens of PBs

[52] https://icannwiki.org/Data_Escrow
[53] https://arstechnica.com/gadgets/2019/11/microsofts-project-silica-offers-robust-thousand-year-storage/
[54] https://www.dpconline.org/events/past-events/webinars

range. This raises questions about the overall performance and possible bottleneck scenarios during operational activities.

The solutions in digital archiving and preservation (please see Table 2) are evolving as the commercial sector matures cloud computing technology and practices for long-term data management. Most of these services offer both backend and frontend solutions. The different types of deployment models offer flexibility for non-profit organisations data management initiatives which are engaging with the market. In order to reap the full benefits, there is a need to continue developing and integrating these solutions in a controlled direction. Most of the general purpose providers of cloud services do not typically address specific data archiving considerations within their basic offerings. Specialist providers have emerged to offer value-added services in archiving and preservation with an extensive list of features available to their users. Many of these companies have focused on simplifying data ingest mechanisms as much as possible in order to speed up their service adoption. This is achieved via automation procedures on content submission, integration with other upstream or downstream solutions, as well as focus on usability with user friendly web-based interfaces with simple "drag-and-drop" features. In some other cases, automation is even extended to approval workflows, metadata parsing, and the embedded preservation processes as part of data management plans. Efforts are taking place in the definition of clear and improved data models which allow the accommodation of new emerging data types such as multi-party emails, social media feeds, or snapshots of large content management systems (CMSs).

However, most of these services have not yet demonstrated mechanisms that allow to easily layer them on top of generic cloud services at scale for research data. There is still a gap between these specialist service providers and generic infrastructure providers. Reasons for this gap may include cost, interoperability, and market segmentation.

Some of these services are following a storage-agnostic strategy which means effort is being put in reducing the aforementioned identified gaps when using generic infrastructure. However, demonstration of full support for portable preservation environments has not yet been achieved and still needs R&D in order to be demonstrated as production quality[55].

Another observation that emerges when analysing commercial digital repository services is the lack in many cases of a designated community[56]. Most of the services offer generalist data repositories services and very few seem to be able to deal with datasets peculiarities and their different usage. Relevant aspects include the scope of data, e.g. coverage, contextual information, provenance and attribution that often do not receive the necessary attention. Some of these repositories seem to have made the choice of entering in the market at an initial stage, even if this means proposing a product that only partially meets the complex needs of those use-cases that require for example more advanced features such as validation and re-use of data.

[55] https://ipres2019.org/static/pdf/iPres2019_paper_111.pdf

[56] https://www.nature.com/sdata/policies/repositories

It remains to be demonstrated how some of the services address issues imposed by the heterogeneity of datasets, native support of data formats, validation procedures offered and the management of data licences.

Other aspects to consider in commercial offers include features such as systematic version control tools, hierarchical relation management (between files, objects and collections) and support for retention policies allowing more advanced data models to be created. This enables a better degree of data curation and metadata handling, but also analytics services to be built on top of the archiving solutions. In addition, it allows as well to establish links with external products using AI (machine learning and deep learning techniques) such as “speech-to-text”, facial recognition, and pattern recognition in photographs.
The creation of monitoring dashboards with real time tracking capabilities of the archive, extended audit trails, and automatic report generation appears to be another significant capability for most of data archiving and preservation service providers.

5.3. Public Sector Archiving and Preservation services

In what concerns in-house services – where they exist – they are being developed by research organisations in order to cover their increasing data preservation needs. Some of them are evolving into service offerings with a growing user base and have started the process of certification schemes such as ISO 16363 or CoreTrustSeal.
In the case of CERN, notable examples include Zenodo, Open Data Portal Service as well as the REANA project.
The Zenodo service is offered by CERN as part of its mission to make available the results in High Energy Physics (HEP). It consists of an open repository hosting a diversity of data from many research disciplines. Zenodo often faces the challenge of accommodating “Big Sciences”, where a given resource may consist of many terabytes of data and associated computational tools are often necessary for interpretation. As with other sciences, uniform data sharing lends itself to aggregation, visualization, and pattern analysis, needing robust sharing and additional advanced capabilities, such as Machine Learning to allow the extraction of patterns or create new data models.
The CERN Open Data portal is a service which disseminates petabytes of primary and derived datasets from particle physics as they are publicly released by the LHC collaborations. The portal offers datasets together with accompanying software examples, virtual machine images, condition databases, configuration files, and necessary associated documentation to enable non-specialists to use the data.
REANA is a project developing a reusable and reproducible research data analysis platform which helps researchers to structure their input data, analysis code, containerised environments and computational workflows, so that the analysis can be instantiated and run anywhere.

To expand the reach of most of these services and make them available at large for the research community, there is a growing interest in exploring the possibility to offer advanced services, such as reproducibility services on top of commercial providers, by instantiating and executing software repositories on public clouds so that research reproducibility can be concretely achieved, independent of any on-premise infrastructure.

In astronomy, mechanisms for the long-term preservation of astronomical data are covered by standards developed, or adopted, by the International Virtual Observatory Alliance (IVOA). The archives for the majority of ESA's astronomy, planetary science and heliophysics missions are developed and maintained by the ESAC Science Data Centre (ESDC)[57], in coordination with the science operations centres, instrument teams and consortia of the various missions. Data from the Gaia, Cluster, Hubble Space Telescope, XMM-Newton, Herschel, ISO, Planck, and many more instruments are already available in the ESDC's archival system.

In life sciences, the best-practice approaches for preserving wet-lab data production show it is important to capture and enable access to specimen cell lines, tissue samples and/or DNA as well as reagents. Wet-lab methods can be captured in electronic laboratory notebooks and reported in the Biosamples database, protocols.io or OpenWetWare; specimens can be lodged in biobanks, culture or museum collections; but the effort involved in enabling full reproducibility remains extensive. Electronic laboratory notebooks are frequently used way to make this information openly available and archived. Some partial solutions exist (e.g. LabTrove, BlogMyData, Benchling and others), including tools for specific domains such as the Scratchpad Virtual Research Environment for natural history research. Other tools can be combined to produce notebooks for small standalone code-based projects, including Jupyter Notebook, Rmarkdown, and Docker. However, it remains a challenge to implement online laboratory notebooks to cover both field/lab work and computer-based work, especially when computer work is extensive, involved and non-modular. Currently, no best-practice guidelines or minimum information standards exist for use of electronic laboratory notebooks.

In Earth sciences, specifically in seismology, EIDA[58] is the European Integrated Data Archive infrastructure within ORFEUS, the Observatories & Research Facilities for European Seismology which is a non-profit foundation to coordinate and promote digital, broadband seismology in the European-Mediterranean area. EIDA is a distributed federation of data centres established to securely archive seismic waveform data and metadata gathered by European research infrastructures and provide transparent access to data for the geosciences research communities. The nine primary nodes signed an MoU with ORFEUS in which they commit to providing continued access to the data in their archives. ORFEUS Data Centre (ODC) is a centralized data centre exploited by ORFEUS to collect and archive waveform data and

[57] https://www.cosmos.esa.int/web/esdc
[58] http://www.orfeus-eu.org/data/eida/

metadata from about 30 European seismic networks. ORFEUS Data Centre is one of the EIDA nodes that offers access to data through EIDA standardized services.

An important fraction of all science is in the long tail of scientific research made up of smaller, less costly projects (LToS). In sciences like geobiology and soil science, for instance, researchers tend to work independently or in small groups, on hypothesis driven research questions, gathering data into privately held collections for local analysis. Currently, these data are rarely shared and re-used, in part because there are no suitable repositories, an issue stemming from the complexity and variation within the practice and culture of small science. Small science data have potential for analysis across aggregates, similar to big science. However, their value for re-use may also be complementary, as a unique piece of a complex puzzle or an important addition to a series of measures over time. The individual researcher might struggle even more to preserve his/her research data. Even if researchers can preserve their data on the infrastructure of the research group or on that of the institute (some organisations do have a clear access policy and infrastructure such as hardware, tools and metadata standards), most often the data is not made available for external access nor is it stored according to FAIR principles.

ARCHIVER has also analysed data archiving services at European level, available in public funded research environments managing the increasing demand to store research datasets. One such service for long-term access and curation of research datasets, with a focus on data from science, engineering and technology present since 2010 is 4TU.ResearchData[59]. It has been managed as a service for researchers (from universities around the world) to deposit and share their data, and for other researchers to download and use data in their research. The Strategy document for the years 2020-2023[60] foresees to follow FAIR principles and recognizes the leading role of EOSC.

The EOSC Executive Board Landscape working group[61] is performing a landscaping analysis of EOSC relevant infrastructures and policies and provides specific information about EOSC relevant e-infrastructures and thematic infrastructures from European (EU) Member States (MS) and Associated Countries (AC). A number of the EOSC relevant infrastructures and policies being documented by the Landscape WG include support for research data preservation and hence are of interest to the work of the ARCHIVER project. The Landscape WG is currently preparing a first draft of its landscape analysis report which is expected to become available in December 2019 and will be taken into account in the planning activities of the ARCHIVER project.

[59] https://researchdata.4tu.nl/en/about-4turesearchdata/organisation/
[60] https://researchdata.4tu.nl/fileadmin/user_upload/Documenten/4TU.ResearchData_Strategy_2020-2023.pdf
[61] https://www.eoscsecretariat.eu/working-groups/landscape-working-group

Still in the EOSC context, the report on "EOSC Hub Technical Architecture and standards roadmap" highlights that data preservation and curation services have received substantial focus in the EOSC-hub plans. B2SAFE is an EOSC-hub service that allows community repositories to implement data management policies on their research data, distributed across multiple administrative domains. B2HANDLE is also an EOSC-hub service and enables users to manage persistent identifiers for data to make them referenceable independently of location, ownership or storage type. It supports metadata associated with identifiers to enable middleware services to execute data management procedures autonomously.

In terms of costs, some funding agencies do not require provisions for long-term data preservation as part of the data management plans requested from grantees. Furthermore, many grant schemes do not consider the costs of data preservation beyond the lifetime of the grant.

The DCC Research Data Management Forum (RDMF) held a meeting in London on 16 September 2019[62] as an effort to clarify costs associated with data management and share knowledge in ensuring appropriate services are in place and required allocations are made. Rudimentary costing tools are available and various pricing plans in practice have been highlighted. Currently, the University of Manchester is reviewing the provision of research data archiving services and how the relevant costs will be assessed. These costs will increase as the University prepares to meet the requirements of funders for making data accessible for minimum periods of time. Currently these costs are covered through overheads. In May 2018, the University decided to change to a Pay Once Store Forever (POSF) model of charging which guarantees an uncharged allowance of TB per funded research project while storage over and above the uncharged allowance will be subject to a once-only charge per TB - this will cover storage for as long as required by funding body regulations, in effect, indefinitely.

This model will be explored in practice by ARCHIVER as one potential business model to satisfy the specific needs of preserving researcher data beyond the lifetime of the projects/grants.

In summary, not all research organisations are offering long-term data preservation services as part of their data management services.

Those organisations that do offer long-term data preservation services tend to propose private or community based in-house solutions in the GB to TB range for individual projects with a freely available limited capacity for all researchers and when possible, additional costs for large scale usage.

A key part of the long-term data preservation services is the local staff expertise in data management offered by the organisations to their researchers. The ARCHIVER project objective is to establish a consolidated process that can take full benefit from the existing support structures for research data management planning combined with the provision of archiving and preservation services by commercial providers.

[62] http://www.dcc.ac.uk/events/research-data-management-forum-rdmf/rdmf19-costing-data-management

At national level, a number of European NRENs are managing country-wide, large-scale, long-term storage services providing a safe storage space to academic users based in their respective countries.[63] These services are based on software platforms funded by government-funded R&D projects and rely on geographically-distributed infrastructure of storage systems (disk arrays, tapes). In some cases, services offer backup and disaster recovery features to secure data in case of own storage breakdown, archiving services of volumes of cold data and cloud storage and synchronization services (e. g. Owncloud). Pilot services are being setup for Long-Term Data Preservation (LTDP) archives, consisting in dark archives with validation of OAIS SIP/AIP packages. Technical measures supported include in some cases (e. g. CESNET DataCare[64]), certified compliance with Information Security Management System Standard EN ISO/IEC 27001:2014.

R&D initiatives not yet close to market also exist in the public sector trying to address the problem of long-term data preservation, in particular to address and higher storage density at infrastructure level.
In the bioinformatics domain for example, R&D for long-term data preservation with biological materials is being performed in order to reduce space and avoid degradation of hardware. In a recent study[65], the Technion-Israel Institute of Technology makes a case for redundancy in DNA storage technologies, achieving storage of information in a density of more than 10 PBs in a single gram of synthetic DNA, whilst significantly improving the data writing process.

In what concerns regulation, the Digital Single Market (DSM) cloud stakeholder group serves as a platform for the work of the two self-regulatory cloud working groups that are currently active: cloud switching/porting data working group and cloud security certification working group.
Under the free flow of non-personal data Regulation[66], the European Commission has encouraged the development of a self-regulatory basis of a Code of Conduct for Cloud Switching/Data Porting to reduce the risk of "vendor lock-in" and create a competitive European digital market where it must be easy to switch from provider, including the porting of data involved. In anticipation of this regulation, SWIPO has been formed as the collective name of a group of stakeholders that are working towards codes of conduct on Infrastructure as a Service (IaaS) and on Software as a service (SaaS) within the domain of Cloud Computing.
Throughout 2018 and 2019 and via a series of meetings and public consultations, the IaaS and SaaS subgroups have been working on their specific codes of conduct that share common details. From the ARCHIVER consortium, CERN is participating in the IaaS subgroup.

63 https://intranet.geant.org/gn4/3/Work-Packages/WP4/Deliverables%20Documents/Community%20Clouds%20Delivery%20Plan%20Services%20Selection,%20Business%20Models%20and%20Development%20Roadmap/D4-1_Community-Clouds-Delivery-Plan-Services-Selection-Business-Models-and-Development-Roadmap.pdf
64 https://du.cesnet.cz/en/start
65 https://www.nature.com/articles/s41587-019-0240-x
66 https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32018R1807

The current planning foresees a release of the Codes of Conduct at the High-Level Conference on the Data Economy, which will take place on 25-26 November in Helsinki.
CERN will share the resulting Codes of Conduct with the ARCHIVER consortium so that they can be taken into account for the execution phase of the project.
The self-regulatory working group on cloud security certification (CSP CERT) aims at exploring options for the development of a possible candidate scheme in the field of cloud security to enhance legal certainty and trust in the cloud market. After 18 months of work, the group presented its final recommendations for a European cloud certification scheme in June 2019 in Amsterdam. They address security requirements, conformity assessment methodologies and assurance levels basic, substantial and high in line with the European Cybersecurity Act.[67]
Security and data protection aspects under the current European legislation (e.g. GDPR) are also key challenges to the existing service providers, especially for companies which have a clear focus on the European market.

5.4. The Challenges to be addressed

ARCHIVER intends to move beyond the existing state of the art, by addressing the identified gaps across the ARCHIVER 4-layer R&D model.
The state-the-art analysis leads to the recognition that archiving and preservation services already exist in the market, however the services currently available have not been proven to scale to the 10s of petabyte range and beyond, nor address the complex unstructured research data types, often domain-specific, that are needed by many scientific disciplines. High, sustained ingest rates and service availability for scientific data from multiple domains have not been demonstrated. This fact raises questions about the performance and specific unknown bottlenecks in the current solutions available, when handling scientific data.
The scalability factor for different types of data is a fundamental part of the ARCHIVER R&D challenge. Therefore, demonstration of functionality of services supporting a data archiving and preservation environment for tens of PBs of scientific data, with high, sustained ingest rates ranging from 1-10 GB/s will be required during the ARCHIVER project.
Another aspect that ARCHIVER will explore in order to address one of the barriers to the adoption of cloud-based services by the research community is the practical implementation of migration and exit strategies. Even if some of the services currently available in the market are following a storage-agnostic strategy, demonstration of full support for migration across preservation environments is still required, with well-established practices, systematic processes and reduced effort[68] both in technical (audits, migrations, virtualisation, etc.) and organisational terms. ARCHIVER intends to push the state of the art in this topic, by probing

67 https://ec.europa.eu/digital-single-market/en/dsm-cloud-stakeholder-working-groups-cloud-switching-and-cloud-security-certification
68 https://ipres2019.org/static/pdf/iPres2019_paper_111.pdf

different implementation of strategies for portability of data preservation environments in order to be able to build viable exit plans and avoid vendor lock-ins, having public funded research organisations acting as data stewards with well-defined data management plans. More specifically, ARCHIVER will explore different deployment model options, so that the resulting software services can be demonstrated to run on top of public clouds, private clouds and in a hybrid on-premises/public cloud model, as an essential part of business continuity and disaster recovery plans, a globally relevant use case for the research community in all the domains.

Another challenge of significant innovation potential to be pursued in ARCHIVER is legal compliance. The level of legal conformity of the solutions in the market is still very variable. The practical implementation challenges that organisations must tackle before they can effectively claim that they can mitigate data privacy related risks can still adversely impact the level of compliance of the current available services. As a result of the OMC, significant R&D is required in order to implement technical and organizational measures in foreseen to comply and follow European regulation (GDPR and recently the Free Flow of Data).

During the R&D phases, contractors will be requested to develop capabilities that will cover aspects such as Data Controller vs. Data Processor roles, mechanisms to record the processing activity when using the ARCHIVER data archiving and preservation services, data retention periods and data transfers.

These aspects will be expanded in the context of applicable codes of conduct and organizational/technical measures foreseen to be developed across the ARCHIVER R&D phases in a "Privacy by Design and by Default" approach, so that the resulting archiving and preservation services can handle correctly and by default personal data present in scientific data sets.

As an essential part of the R&D to be performed, ARCHIVER will request Contractors to develop innovative business models for research data repositories. On one hand, there is still a gap to be addressed in stimulating cross-disciplinarily of archiving and preservation of scientific data, specifically by harmonizing, planning and integrating the costs associated with the operation of archiving and preservation services. ARCHIVER will explore for example, models that can allow organisations or individual researchers to store their data after the end of their research grant. On the other, the current business models for cloud-based services are not particularly adapted to data archiving and preservation needs with a timeline of decades. In particular, this aspect is more evident when offering some degree of protection against changes in the market is required, in scenarios such as supplier bankruptcy. ARCHIVER will also explore the voucher concept, a practice pioneered for the scientific community in HNSciCloud[69]. ARCHIVER will build on those results and lessons learned exploring it as an innovative business model and incentive to turn scientific data FAIR.

The Table 2, summarizes the R&D the ARCHIVER project intends to perform in order to move the current state of the art in the desired direction.

[69] https://zenodo.org/record/2615456#.XcMgc0VKhxk

Layer	Current State-of-the-Art	ARCHIVER R&D	
Layer 1 Storage Basic Archive Secure backup	Deployment over private, public, hybrid, community and special purpose Clouds in a range of single PBs.	Infrastructure agnostic archiving and preservation services with performance demonstration in the tens of PBs of scientific data volume with linear growth over the years. Demonstration of support of multiple tenancy with data and access isolated. Sustained data ingest rates capabilities from 1-10 GB/s. Connectivity to the GEANT network. Follow best practices as foreseen in ISO 27000/27040/19086 and European Cybersecurity Act.	Take into account practices as foreseen in the SWIPO Codes of Conduct. Demonstration of Implementation of viable exit strategies across providers. Demonstration of a "Privacy by Design" approach, with implementation of technical and organisational measures in order to protect personal data under GDPR. Implementation and demonstration of innovative business models, planning and integrating cost of long-term data archiving, allowing, for example, organisations or individual researchers to store their data after the end of their research grant. Explore the voucher concept as an innovative business model and incentive to turn scientific data FAIR. Reduce procurement time for commercial archiving and preservation services by 50%.
Layer 2 Preservation	Preservation of files, folders, Content Management Packages (archives) and data types. Emails backup in pdf format. Migration and backup plans, APIs, SFTPs, Web Interfaces, Drag&Drop ingestions. Metadata and context processing.	Preservation for more than a decade in mind. Follow best practices foreseen in OAIS, ISO 14721/16393, 2632, CoreTrustSeal in terms of self-assessment. Support for handling of unstructured and missing metadata. Co-creation and assessment of data management plans in collaboration, across data producers and commercial repository services. Testing and implementation of clear models for local support responsibilities for long-term data management planning at the Buyer organisations. Use of open source software and vendor independent standards and interfaces (such as PREMIS, METS and Bagit) to allow exit strategies and prevent vendor lock-ins. Provision of generic open APIs, in order to be able to integrate the resulting	

		services in innovative workflows such as moving data according to its access "temperature" in/out of long-term archives or repositories.	
Layer 3 User services	Volumes of hundreds of TBs with support of Indexing, elastic search, deduplication, single point access, crawling, cross-checking, vulnerability scanning, plugins configuration.	Access and permission management against repositories and various collections supporting Federated Identity and Access Management schemes in the research community. Support of automated indexing of of tens of PB of content for fast information tagging, easy and broader search as a strategy to promote open data access.	
Layer 4 Advanced services	Support of retention and integrity of certain types of data over a decade needed ensuring it is tamper-proof whilst allowing easy access and basic re-usability.	Full reproducibility of services (initial examples are database services and/or software distribution services) on top of the resulting archives. 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. (iii) Support open data analyses run directly by researchers in the resulting services, using containerised environments residing in the external commercial service provider infrastructure, ensuring full reusability of open data, software and the compute/storage environment, independently of the original on-prem infrastructure.	

Table 2: ARCHIVER R&D foresseen across the 4-layer architecture.

As part of the challenges in R&D, ARCHIVER will also consider throughout the project the environmental impact. As part of the efforts to reduce the carbon-footprint, the EC launched a tender in 2018 to produce a study on "Energy-efficient Cloud Computing Technologies and Policies for an Eco-friendly Cloud Market"[70] with a focus on cloud computing technologies. It aims to develop a set of sound recommendations for energy-efficient digital services in the cloud domain that will contribute to the goals of the Paris Agreement. It is also investigating policy options for driving the market towards procuring and offering energy-efficient ICT services. The study is currently ongoing with its current actions and developments being published on the study's website[71]. ARCHIVER will try to implement as possible during the call-off transition phases, the recommendations to be produced by the study concerning procurement of energy-efficient cloud based services by public organisations.

6. Conclusions

The Open Market Consultation process, including the state-of-the-art analysis and community landscaping as an integral part of it, were crucial activities for identifying the demand side requirements and assessing the capacity from the supply to meet them. These are summarized in section 2 and 3 of this document.

The dialogue with the supply-side during the OMC showed that some functionalities embedded in a layered model require significant and realistic R&D efforts from industry.

[70] https://ec.europa.eu/digital-single-market/en/news/energy-efficient-cloud-computing-and-green-digital-services

[71] https://www.cloudefficiency.eu/home

These challenges can be resumed as follows:

- Cost-effective deployment models
- Data Privacy and Sovereignty
- Best Practices (OAIS Reference Model and CoreTrustSeal) conformance
- FAIR Principles
- Management of scientific datasets beyond bit preservation
- Network Connectivity & Federated AAI
- Scalability from Terabyte to Petabyte

The landscaping activities showed the importance of aligning and integrating services making them broadly available in order to reduce fragmentation, based on standards, adherent to European legislation with transparent business models.

In particular, the deployment models' flexibility of the resulting solutions will be very important as the lack of verified data exit plans remains one of the biggest factors hampering the adoption of cloud based preservation services by research organisations, raising concerns of vendor lock-in and inability to migrate across services.

ARCHIVER resulting services will be very pertinent for the European landscape of data preservation.
The project results aim to concretely implement a combination of best practices for research data management across different disciplines with a set of resulting archiving and preservation services, based on trustworthy repositories, where sustainability of the funding model is a fundamental factor. These services aim to be fully integrated in the EOSC, following the motto *"Data and Services made in Europe"* being trustworthy and transparent, aligned with the vision of creating an integrated environment across public and private sectors, science producers and science consumers.
The core objective is to stimulate new ways of making research resulting in real innovation across disciplines and geographies. As the rules of engagement with the EOSC evolve, specifically for the private sector, they will be presented to the contractors engaged in the project.

The overall outcome of the OMC activities is reflected in the PCP Contract Notice. Specifically, the selection and award criteria address the R&D challenges to ensure that the selected bids are capable of meeting the demand-side requirements.

7. Lessons learned and considerations for future Open Market Consultations

This section contains some of the lessons learned during the Open Market Consultation and considerations for the organisation of similar Consultations in the context of future projects.

The iterative and transparent drafting process of the PCP Contract Notice has brought value and significantly improved the quality of the tender specifications. The comments from demand-side and supply-side stakeholders on the documents allowed the identification of loopholes and ambiguities and helped to fine-tune the requirements. The iterative process allowed potential tenderers to be as informed as possible, and to raise their awareness about the requirements at a very early stage of the dialogue. Consequently, it gave potential tenderers the opportunity to ask for clarifications during the OMC events, providing sufficient time to prepare for the tender and to build solid consortia.

The ARCHIVER OMC was designed as an evolutionary process comprising several events focusing on different user communities, requirements and aspects of the procurement process. Four events targeted to potential tenderers and open to end-users took place in the country of the Buyer organising the event. This had the very positive effect of easing the participation of end-users from different communities and potential tenderers from different regions while reducing the environmental impact of transport. Participating firms could also participate remotely via webcast to several of the events.
For the Barcelona and Stansted events, the project team liaised with local organisations such as Enterprise Europe Network (EEN) to disseminate and promote ARCHIVER to local and regional stakeholders. The use of these organisations, specifically oriented to increase competitiveness of SMEs, has proven to be valuable. The project team witnessed an increase in participation of local SMEs at EEN promoted workshops.

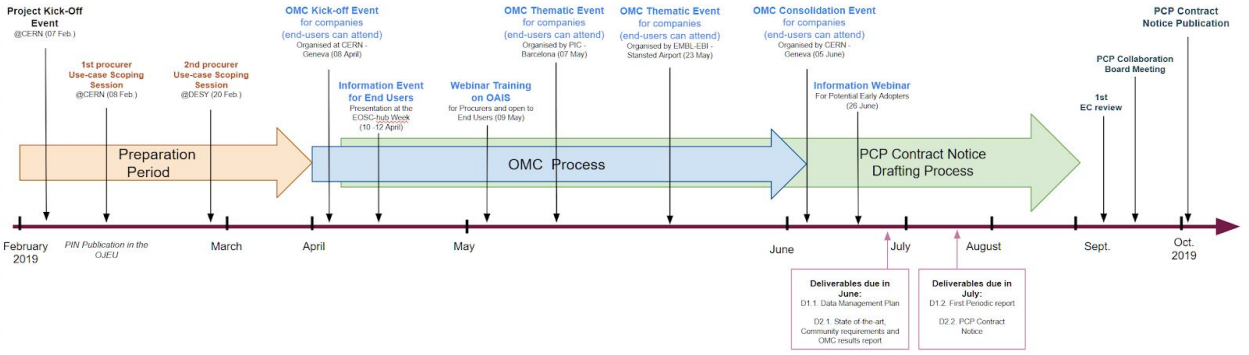
The use of a moderation method, such as the planning poker, helped scope the discussions around the demand-side requirements and their risks for the supply-side. To ensure the objectivity of the moderator, this role should not be carried out by one of the Buyers but rather by a third party organisation. However, the method would be more efficient if the dialogue starts with a clear overview of the R&D challenge illustrated by real-life deployment examples before deep diving into atomic use cases underpinning the general solution.

When it comes to the choice of the organisation performing this task, it should be driven both by the expertise in moderation techniques and technical knowledge on the procurement domain.

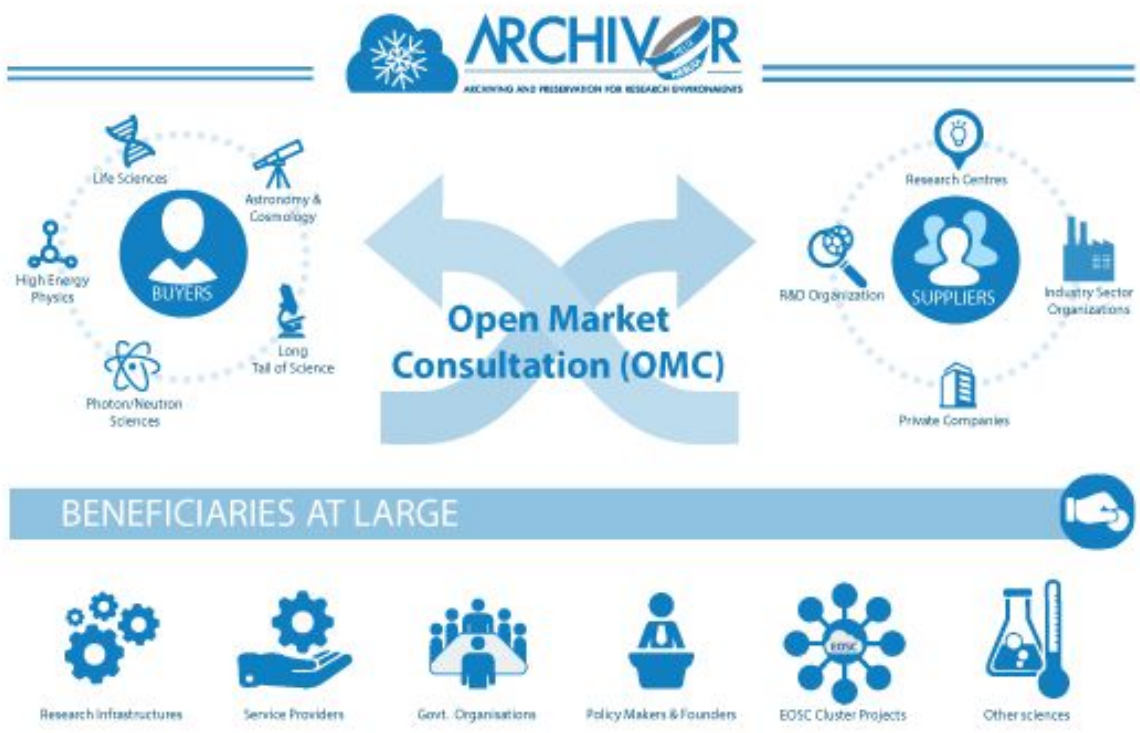
A range of skill sets are needed to achieve the R&D results of a PCP project. The quality of the bids received can be considerably improved if companies combine their skills and resources to form viable consortia to respond to the tender. The project team can encourage the formation of consortia by making available tools to facilitate the partner search.
In ARCHIVER, the choice to use the Crowdhelix platform has proven to be useful. However, companies participating in the ARCHIVER OMC reported that an active moderator role in the formation of consortium would be of great value. ARCHIVER therefore recommends subcontracting an organisation to moderate the formation of consortia in view of the tender. Moderation also requires knowledge of the procurement domain.

Finally, assessing the current state-of-the-art of potential tenderers is a highly important task to scope the project R&D. This activity can take place also during the Open Market Consultation where potential tenderers can be invited to present their state-of-the-art solutions to the project consortium, the end-users, potential early-adopters and even to other companies that can potentially become partners in a consortium. ARCHIVER has invited potential bidders to share their current state-of-the-art via recorded webinars, that subject to their approval, have been made publicly available on the ARCHIVER website.

Appendix A: Open Market Consultation Timeline



Appendix B: ARCHIVER Stakeholder at a glance



Appendix C: Participation in the Digital Preservation Webinar Training

Total number of participants	**18**

Organisation's Name	Role in ARCHIVER	Number of participants
PIC	Consortium Member	4
CERN	Consortium Member	5
DESY	Consortium Member	2
EMBL-EBI	Consortium Member	1
ESA	Potential Early Adopter	1

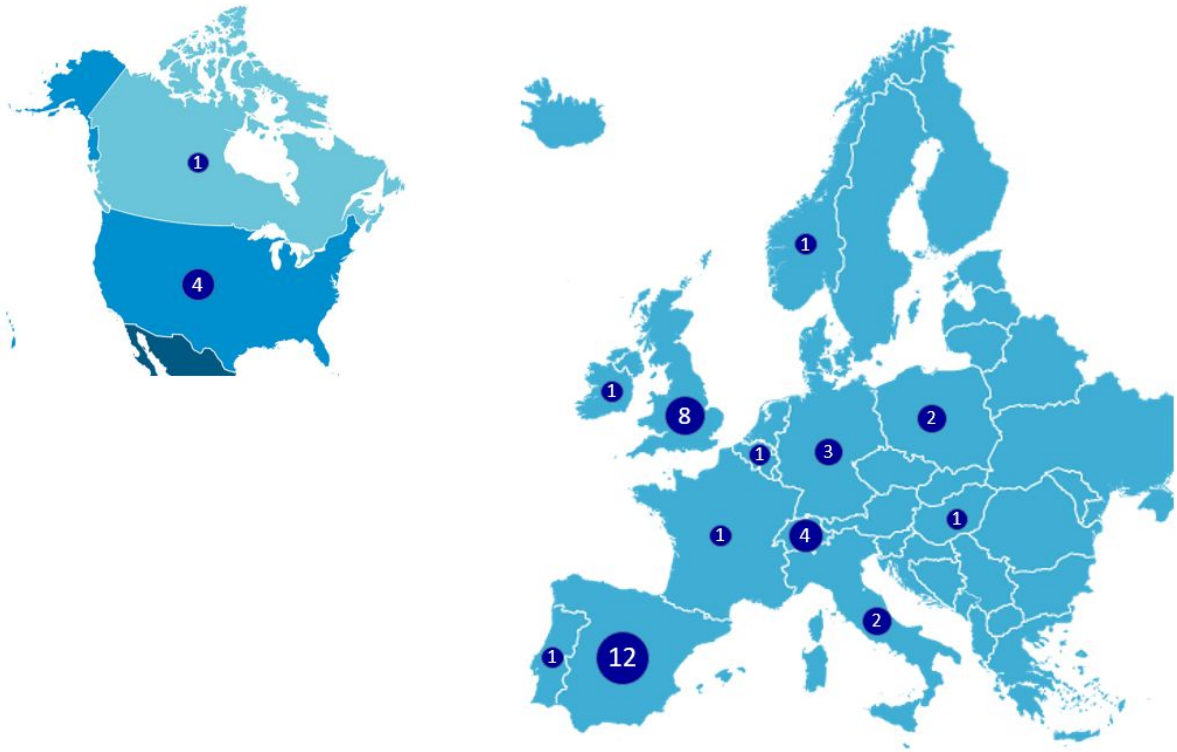
Company's Name	Field of Activity	Number of participants
Gmv	System Integrator / Information Security	2
Dedagroup	Digital Preservation / Archiving Company	2
Libnova	Digital Preservation / Archiving Company	1

Appendix D: Atomic Use Cases

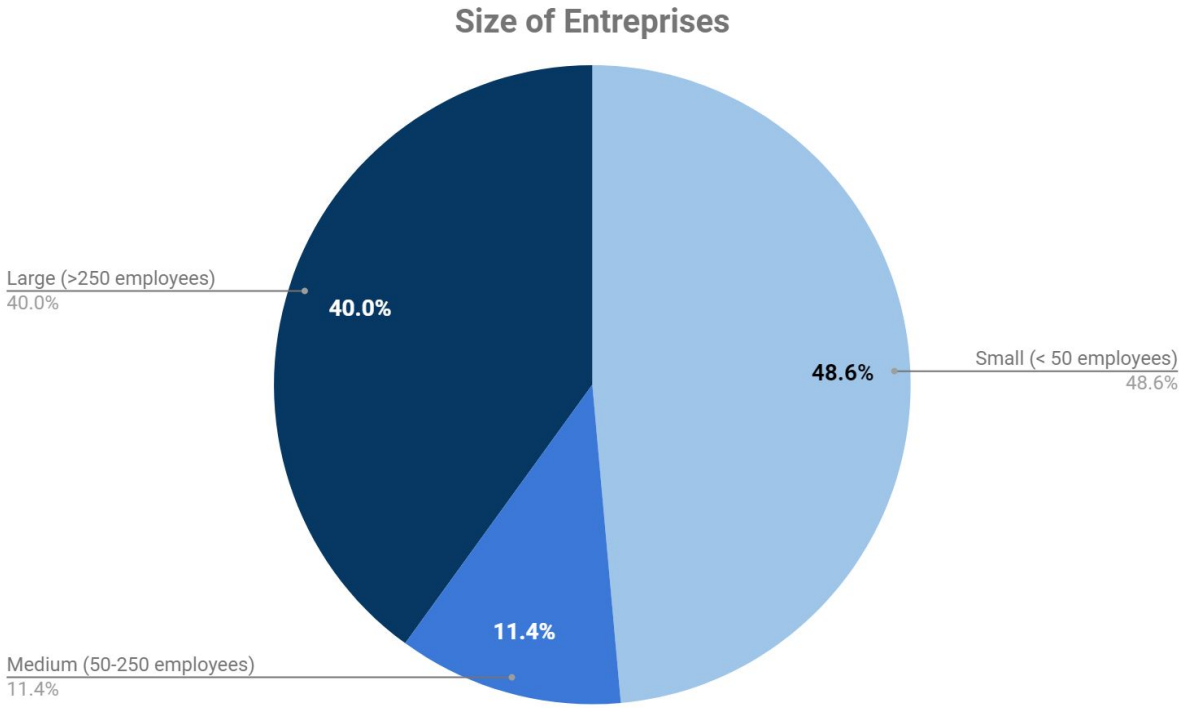
Appendix E: Attendance at the Open Market Consultation Events

Event	Number of Participants Physically	Number of Participants Remotely	Number of Companies	Number of Public Institutions
Kick-off Event (Geneva)	70	42	28	1
Barcelona Event	32	8	16	1
Stansted Event	40	NA	14	1
Consolidation Event (Geneva)	30	23	17	3
Total	**245**		**75**	**6**
Number of different companies / organisations	NA	NA	**38**	**4**

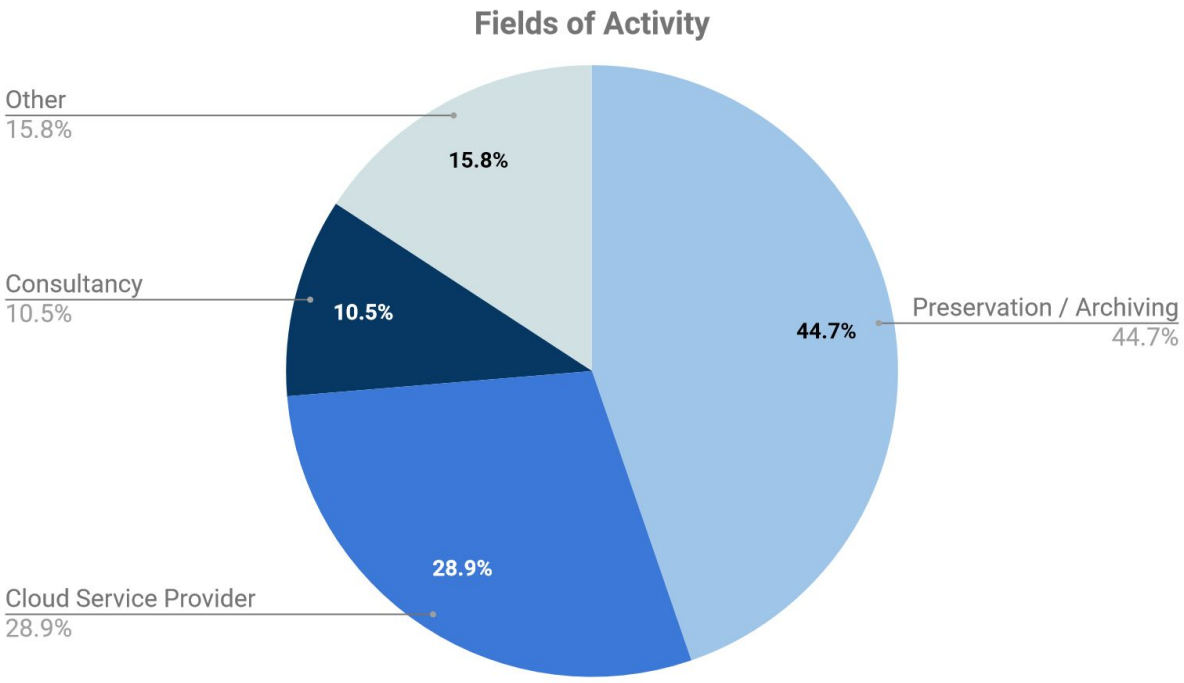
Appendix F: Geographical Distribution of the Organisations attending the OMC events



Appendix G: Size of the Enterprises attending the OMC events



Appendix H: Fields of Activity of the Entreprises attending the OMC events



Appendix I: List of Companies and Organisation that participated in the Open Market Consultation Events

Company's Name	Location
Data Preservation / Archiving Company	
Arkivum	UK
Artefactual Systems Inc	Canada
Atempo	France
Ddoceo Software	Spain
Dedagroup	Italy
Docuteam GmbH	Switzerland
Dropbox International	Ireland
Figshare	UK
GRAU DATA GmbH	Germany
KEEP SOLUTIONS	Portugal
Libnova	Spain
Piql	Norway
Preservica	UK
Proofpoint Inc.	United States
Rhea GROUP	Belgium
Scipedia, SL	Spain
TECNIO Centre EASY	Spain
Cloud Services Providers	

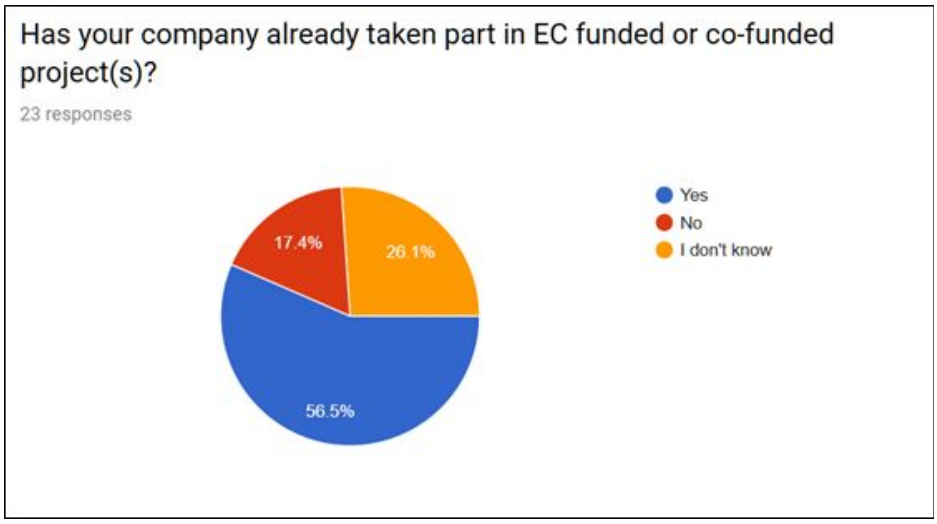
Amazon Web Services (AWS)	United States
BIOS IT	UK
CloudFerro sp. z O.O.	Poland
Exoscale	Switzerland
Huawei	Germany
IBM	UK
Onedata	Poland
Oracle	United States
Tata Consultancy Services Switzerland Ltd	Switzerland
Tessella	UK
T-Systems International GmbH	Germany
Consulting Companies	
AdviceNET	Hungary
Dribia Data Research	Spain
Everis Spain	Spain
JUSTINMIND	United States
Others	
Crowdhelix	UK
Flashman Fotografia	Spain
GMV Secure e-Solutions	Spain
Safelayer	Spain
Submer	Spain
TELECOM ITALIA SPARKLE	Italy

List of public organisations participating in the OMC events

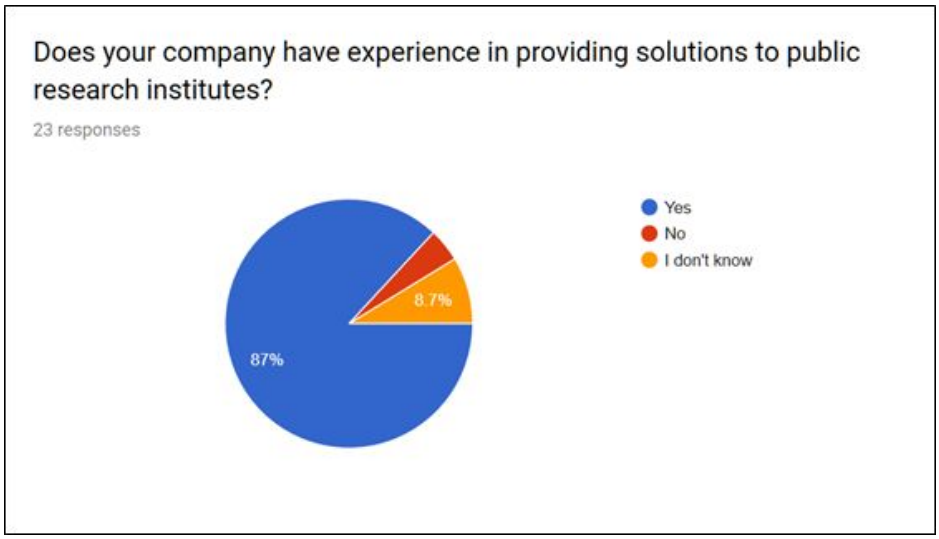
Organisation's Name	Organisation's Type	Location
CELLS- ALBA Synchrotron	Research Institution	Spain
Jisc	NREN	UK
SWITCH	NREN	Switzerland
TECNIO Centre EASY	Research Institution	Spain

Appendix J: Summary of the feedback from the OMC events

Experience

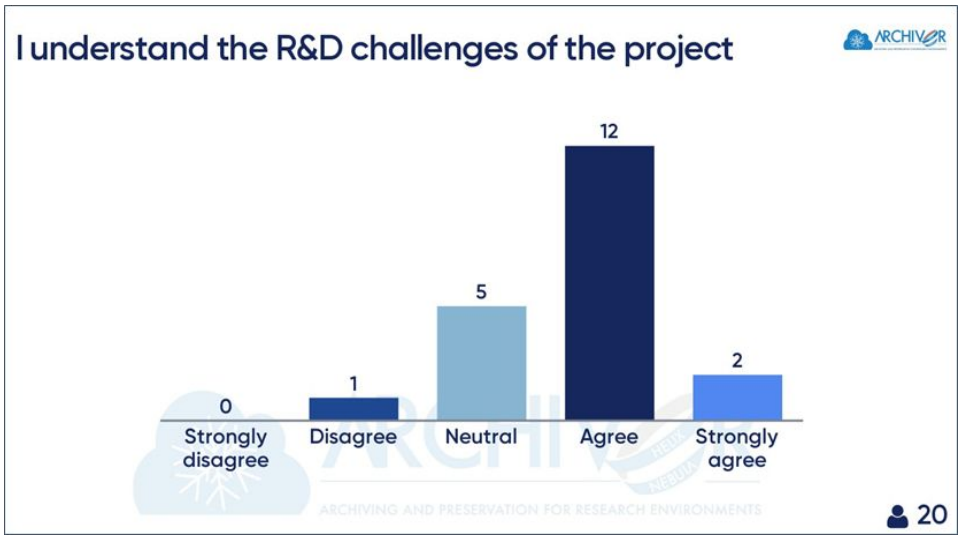


Feedback received at the Kick-off event

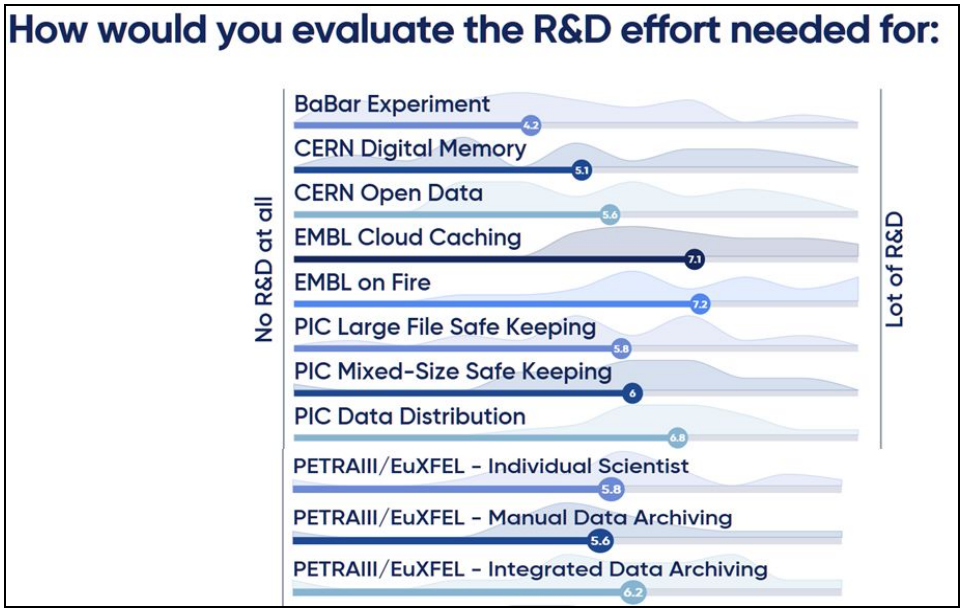


Feedback received at the Kick-off event

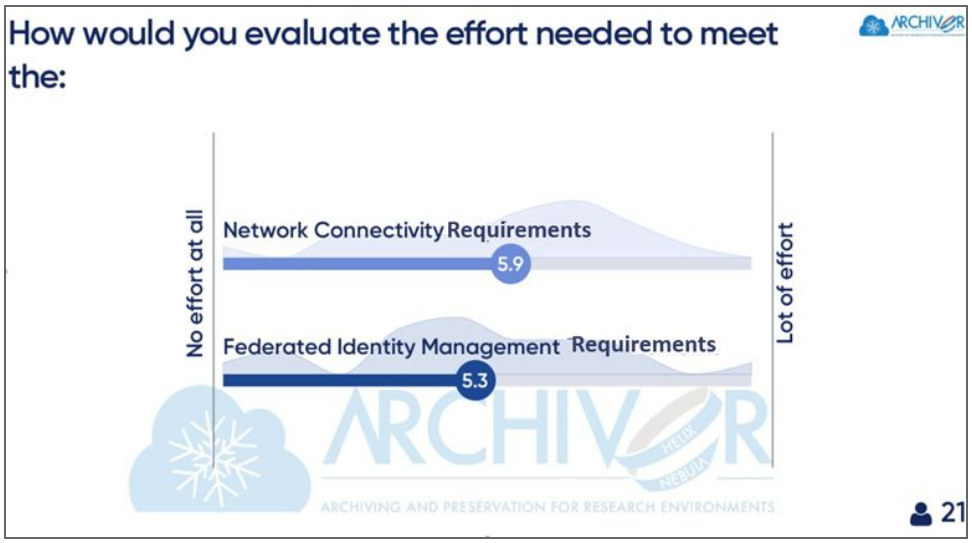
Assessment of the R&D challenges



Feedback received at the Barcelona event

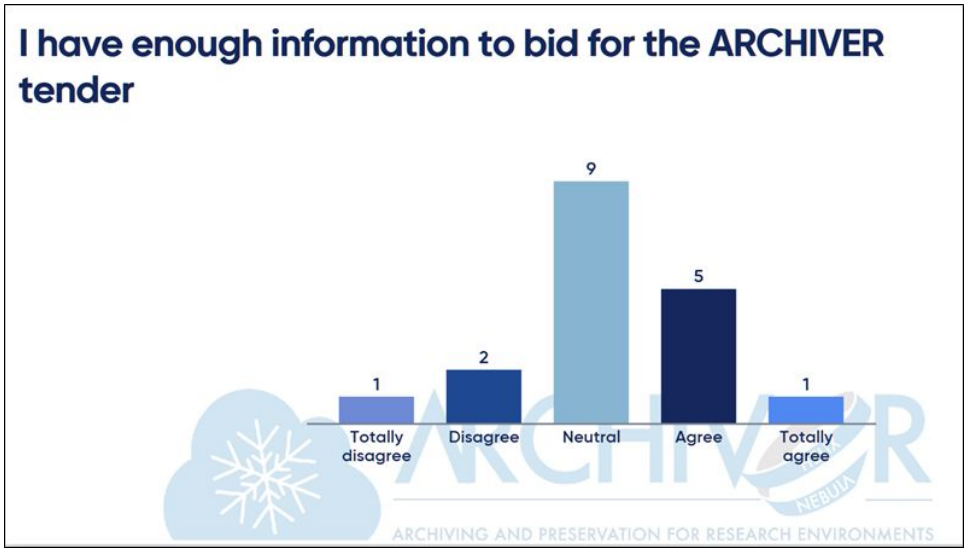


Feedback received at the Stansted and the Consolidation events

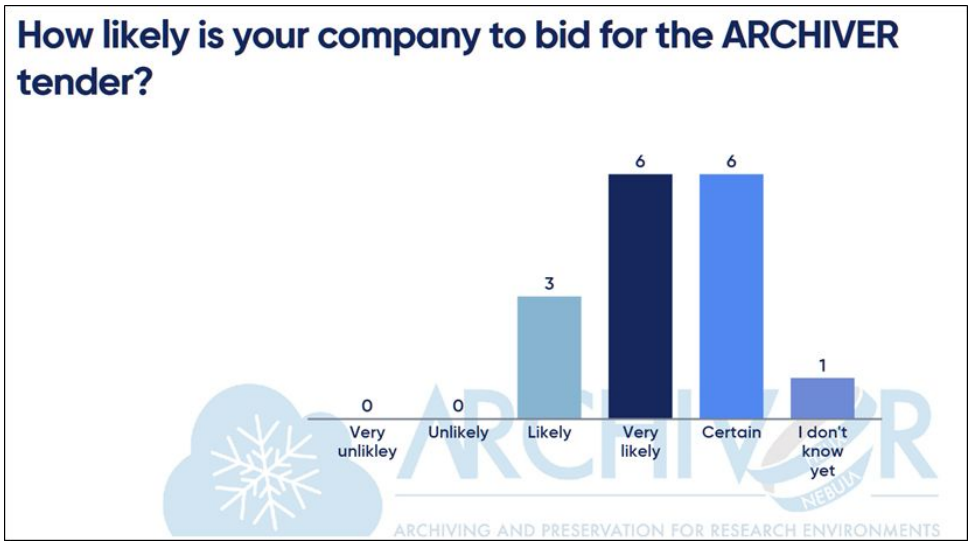


Feedback received at the Stansted event

Intentions:

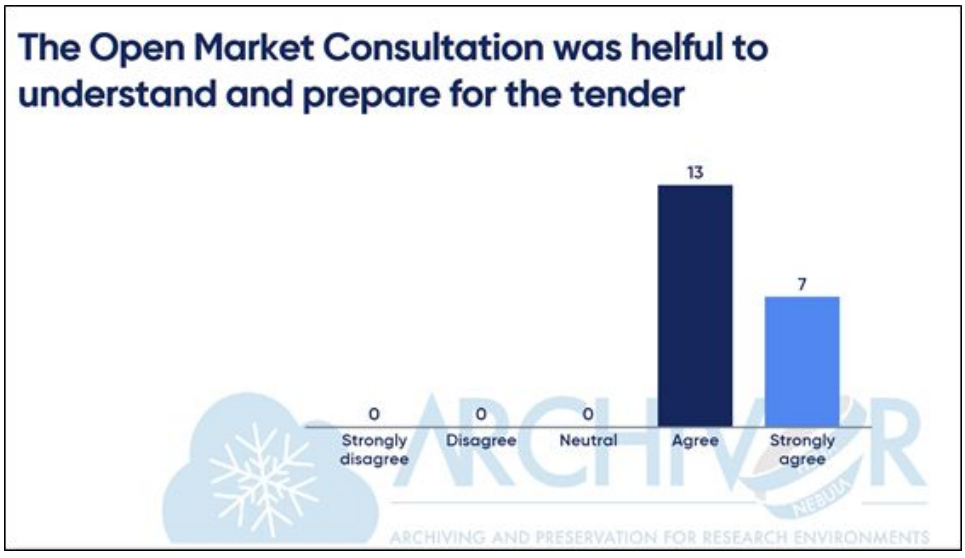


Feedback received at the Consolidation event

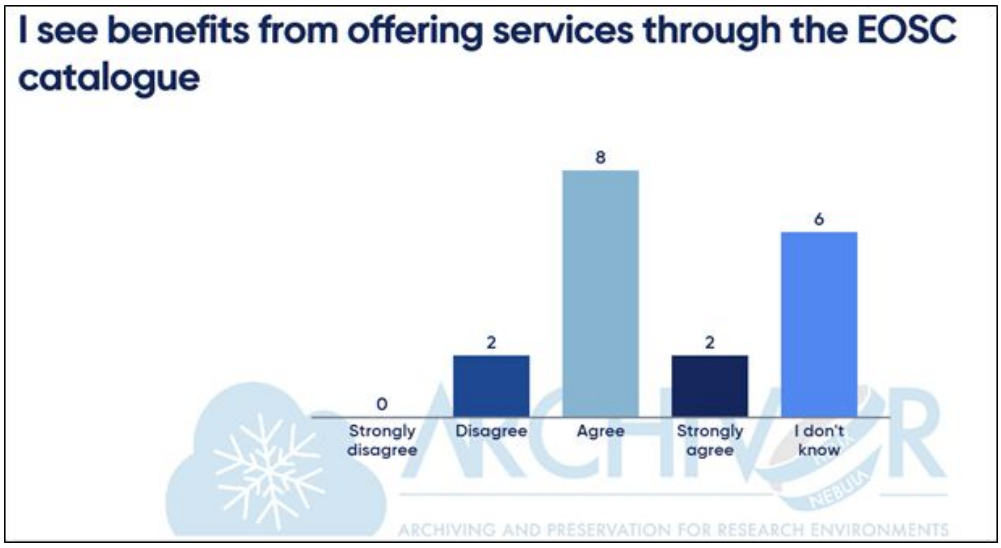


Feedback received at the Consolidation event

General Feedback:



Feedback received at the Consolidation event



Feedback received at the Consolidation event