PROTOTYPE PHASE KICK-OFF EVENT AND AWARD CEREMONY

07 December 2020

Contact: info@archiver-project.eu
Project website: www.archiver-project.eu
Event Outline

14:00 - 14:10: Welcome from Tony Wildish (EMBL-EBI)
14:10 - 14:20: Project overview / update - Joao Fernandes (CERN)
14:20 - 15:00: Expected outcomes of the Prototype Phase - Buyers Group representatives (CERN, DESY, EMBL-EBI, PIC)
15:00 - 15.10: Early Adopters Programme - Anna Manou (CERN)

15:10 - 15:20: Break

Award ceremony (by reverse alphabetical order):
15:20 - 14:35: Presentation from T-Systems International, GWDG and Onedata
15:35 - 15:50: Presentation from LIBNOVA, CSIC, University of Barcelona, Giaretta Associates, AWS and Voxility
15:50 - 16:05: Presentation from Arkivum and Google Cloud
16:05 - 16:20: Closing remarks & Mentimeter - Marion Devouassou (CERN)
Welcome!

Prototype Phase Public Awards Ceremony
December 7th 2020

Tony Wildish – EMBL-EBI
What is EMBL-EBI?

- Europe’s home for biological data services, research and training
- A trusted data provider for the life sciences
- Part of the European Molecular Biology Laboratory, an intergovernmental research organisation
- International: 650 members of staff from 66 nations
EMBL member states

Austria, Belgium, Croatia, Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Luxembourg, Malta, Montenegro, the Netherlands, Norway, Portugal, Slovakia, Spain, Sweden, Switzerland and the United Kingdom

Associate member states: Argentina, Australia

Prospect member states: Lithuania, Poland
Our mission

Deliver excellent research
Deliver scientific services
Train the next generation of scientists
Engage with industry
Coordinate bioinformatics in Europe
The European Molecular Biology Laboratory

80+ nationalities

Heidelberg, Germany
Main Laboratory
Tissue Biology, Disease Modeling
Barcelona, Spain

>1700 personnel

Hinxton, Cambridge, UK
Bioinformatics
Mouse Biology
Rome, Italy

6 sites in Europe

Grenoble, France
Structural Biology

Hamburg, Germany
Structural Biology
Data resources at EMBL-EBI
Database interactions

- Data exchange between EBI data resources

- Arc width weighted by the number of different data types exchanged
Data volume doubles every two years
• => half of our data will always be < 2 years old

EGA and ENA account for the bulk of the data
• DNA sequences

BioImaging repository
• Just starting, will be big
And is getting cheaper to produce

Our data comes from everywhere
Who uses EMBL-EBI services?

See the live map at www.ebi.ac.uk/about/our-impact
Video: Outcomes of the design phase
Project overview / update

João Fernandes – CERN
Focus: Archiving and Data Preservation Services using cloud services available via the European Open Science Cloud (EOSC)
Procurement R&D budget: 3.4M euro; Total Budget: 4.8M
Starting Date: 1\textsuperscript{st} of January 2019
Duration: 42 Months
Coordinator: CERN (Lead Procurer)
Consortium

Includes Buyers and Experts in the preparation, execution and promotion of the procurement of R&D services

The “Buyers Group”: Public organisations committing funds to contribute to a joint-R&D-procurement, research data use cases and R&D testing effort

Experts – Partner organisations bringing expertise in requirement assessment and promotion activities, not part of the Buyers Group
Progress beyond the state of the art

<table>
<thead>
<tr>
<th>Current Scientific Data Repositories</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Growing data volumes</strong></td>
</tr>
<tr>
<td><strong>Basic bit preservation capabilities</strong></td>
</tr>
<tr>
<td><strong>Most of research data not available</strong></td>
</tr>
<tr>
<td><strong>Technology lock-ins concerns (tape), Business Continuity plans needed (COVID-19)</strong></td>
</tr>
<tr>
<td><strong>Fragmentation across scientific disciplines &amp; countries</strong></td>
</tr>
<tr>
<td><strong>Cost underestimation at the planning phase</strong></td>
</tr>
</tbody>
</table>

ARCHIVER “current state of the art” report: https://doi.org/10.5281/zenodo.3618215
## R&D Scope

### Demand Side Requirements

<table>
<thead>
<tr>
<th>Layer 4</th>
<th>Advanced services</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>High level services: visual representation of data (domain specific), reproducibility of scientific analyses, etc.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Layer 3</th>
<th>Baseline user services</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>User services: search, discover, share, indexing, data removal, etc. Access under Federated IAM</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Layer 2</th>
<th>Preservation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OAIS conformant services: data readability formats, normalization, obsolesce monitoring, files fixity, authenticity checks, etc. ISO 14721/16363, 26324 and related standards</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Layer 1</th>
<th>Storage/Basic Archiving/Secure backup</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Data integrity/security; cloud/hybrid deployment</td>
</tr>
<tr>
<td></td>
<td>Data volume in the PB range; high, sustained ingest data rates. ISO certification: 27000, 27040, 19086 and related standards. Archives connected to the GEANT network</td>
</tr>
</tbody>
</table>

---

Scientific use cases deployments documented at: [https://www.archiver-project.eu/deployment-scenarios](https://www.archiver-project.eu/deployment-scenarios)
ARCHIVER & ESFRI science cluster projects

ESCAPE
European Science Cluster of Astronomy & Particle physics ESFRI research Infrastructures

EOSC-Life

panosoc
photon and neutron open science cloud

SSHOC
social sciences & humanities open cloud

ENVRI
LIFE IS NOT FAIR OUR DATA IS FAIR
WWW.ENVRI.EU
Early Adopters  [https://archiver-project.eu/early-adopters-programme](https://archiver-project.eu/early-adopters-programme)

- **Participants:**
  - Demand side public sector organisations
- **Key advantages**
  - Assess if resulting services address archiving and preservation meet their needs
  - Contribute and shape the R&D carried out in the project, contribute with use cases
  - Have the option to purchase pilot-scale services by the end of the project
- **Confirmed 12 organisations**

[Images of logos from CRG, IEO, SURF, SARA, aarnet, SWITCH, KAUST, STU, UKRI, Science and Technology Facilities Council, and Jisc]
“We are creating a European Open Science Cloud now. It is a trusted space for researchers to store their data and to access data from researchers from all other disciplines. We will create a pool of interlinked information, a ‘web of research data’. Every researcher will be able to better use not only their own data, but also those of others. They will thus come to new insights, new findings and new solutions.”

Ursula von der Leyen, European Commission President
World Economic Forum in Davos, January 2020

slide courtesy of Bob Jones (EOSC Sustainability Working Group, CERN)
The Vision

Building the EOSC ecosystem collaboratively with all stakeholders through the EOSC Partnership

- Enable interdisciplinary research to address societal challenges
- Support Open Science and contribute to the Digital Single Market
- Offer EU researchers the digital resources they need to practise Open Science
- Reduce fragmentation by federating existing research infrastructures
- Develop a Web of FAIR Data and Services (including publications and software)
- Give Europe a global lead in research data management
- Stimulate the emergence of a competitive EU cloud sector

slide courtesy of Bob Jones (EOSC Sustainability Working Group, CERN)
First iteration – a Minimum Viable EOSC (MVE)

- Establish an initial MVE that will enable the federation of existing and planned research data infrastructures

- MVE includes EOSC-Core and EOSC-Exchange which work with federated FAIR datasets

- Main focus and added value: connect disciplinary infrastructures and enable cross-disciplinary research
• Broad pan-European requirement analysis of the research sector
  • Analysis results considered in the competitive R&D tender
  • Technical and organisational measures aligned with European legislation in the services being developed (by default & by design)

• Early Adopters Programme established
  • Additional use cases expanding further the set of supported scientific domains
  • Publicly funded research actors external to the ARCHIVER consortium

• Model to facilitate procurement of sustainable pilot services
  • For consortium members and Early Adopter organisations
  • Beyond the lifetime of the project

ARCHIVER is the only EOSC related H2020 project focusing on Archiving & Long Term Data Preservation services for PetaByte scale datasets across multiple research domains and countries.
Design Phase Highlights

• The objectives of the design phase were successfully met by all 5 consortia.

• Main R&D challenge and scientific use cases requirements were globally understood.

• CERN, EMBL-EBI, DESY & PIC allocated significant effort assessing and testing the demo platforms, ingesting data, showcasing current capabilities and state-of-the-art.

• Continuous dialog between research performing organisations and service providers.

• The project team was congratulated for the excellent interaction, generating good progress when compared to other project formats, including project dissemination actions.
Selected Consortia for Prototype Phase

- T·Systems
- Open Telekom Cloud
- One DATA
- GWGDG
- Libnova
- CSIC
- Universitat de Barcelona
- Giaretta Associates
- Google Cloud
- Arkivum
- Bringing archived data to life

https://www.archiver-project.eu/
https://twitter.com/ArchiverProject
https://www.linkedin.com/company/archiver-project/
https://www.youtube.com/channel/UCCBIyLPUt-hWmQatgdlhIzw
Expected outcomes of the Prototype Phase

Buyers Group representatives

CERN, DESY, EMBL-EBI, PIC
CERN Requirements and Expectations

Jean-Yves Le Meur, Tibor Simko, Jakub Urban
CERN Open Data: rich preservation

Simulated dataset QCD_Pt_170_250_EMEnriched_TuneZ2star_8TeV_pythia6 in AODSIM format for 2012 collision data
/QCD_Pt_170_250_EMEnriched_TuneZ2star_8TeV_pythia6/Summer12_GSR3X_PU_RD1_START53_V7N-v1/AODSIM/CMS collaboration

Cite as: CMS collaboration (2017). Simulated dataset QCD_Pt_170_250_EMEnriched_TuneZ2star_8TeV_pythia6 in AODSIM format for 2012 collision data. CERN Open Data Portal. DOI:10.7483/OPENDATA.CMS.ZT85EJNY

Description
Simulated dataset QCD_Pt_170_250_EMEnriched_TuneZ2star_8TeV_pythia6 in AODSIM format for 2012 collision data.

See the description of the simulated dataset names in: About CMS simulated dataset names.

These simulated datasets correspond to the collision data collected by the CMS experiment in 2012.

Dataset characteristics
30132626 events, 269958 files, 9.6 TB in total.

“Bags of bags” for archiving complex datasets
(example: 1 dataset, 26K files, 9.6 TB)

Multiple checksumming options
Hierarchical data organisation

https://www.archiver-project.eu/deployment-scenarios-technical-summaries/cern-open-data
CERN Open Data: towards reuse and reproducibility

- $ file mycode1.cc
  mycode1.cc: C source, ASCII text
- $ file mycode2.cc
  mycode2.cc: Python script, ASCII text executable
- $ file mydata.csv
  mydata.csv: CSV text
- $ csivolint mydata.csv
  Record #15 has error: wrong number of fields in line

Community-oriented data exposure

https://www.archiver-project.eu/deployment-scenarios-technicalsummaries/cern-open-data
CERN Digital Memory: Archive for institutional data

Diversity of input types: text, image, videos in many formats

Multiple pipelines for different sources

Challenges to address: duplication, authorships, integrity, versioning
CERN Digital Memory: AICs & AIUs

AIC versioning strategy

- Metadata-only & File-only AIPs required: one record / many files
- Naming convention of submitted objects with relevant IDs
- Support for BagIt structure with reference files
- Ability to reconstruct original objects at any given point of time

CERN OAIS Registry update with AIC IDs, possibly using webhook calls
DESY Requirements and Expectations

Sergey Yakubov, Martin Gasthuber
Main sources of data to be archived and preserved

- two sites
  - Hamburg
  - Zeuthen (near Berlin)
- science areas
  - particle physics (LHC, Belle 2, ...)
  - photon science (EuXFEL, Petra III, FLASH)
  - accelerator research (wakefield, Petra IV, ...)
  - astrophysics
- all areas “data intensive science”
**individual scientist / small working groups**
- scientist is the archivist
- publication material + condensed data + reference to full datasets
- DOI handling
- mainly interactive access
- few TB, 100MB/sec, 10K objects
- ~0.2-0.5PB annual
- more or less ‘classical preservation model/practices’

**mid-size working groups (Petra III experiment)**
- nominated member of the group is the archivist (on behalf of)
- raw + derived data + code
- DOI + open-data handling
- comply with site data policy
- few 10TB, 1-2GB/sec, >150K objects
- <50% interactive access
- ~2-4PB annual

**large collaboration / site management (EuXFEL organization)**
- site nominated archivist responsible for all experiments
- raw + calibration data + code
- DOI + open-data handling
- comply with site data policy
- few 100TB, 2-10GB/sec, >30K obj.
- very low interactive access
- >30PB annual
More concretely / general expectations

- functionalities and features to be completed in this phase
  - last minute changes early next phase

- full focus on scaling and stability (at the same time ;-) at next phase
Case I - small size - Individual Scientist

- Simple & Small - very similar to classic data preservation use cases
- accessed mainly via a web browser (GUI) from single user
- extras / probably not covered by existing solutions
  - authentication - binding to local IdP
  - metadata scheme - added community specifics on top of standards
  - DOIs
  - local & hybrid deployments (data preservation core, metadata core, storage)

- just do it
Case II - mid size - Petra III Experiment

- Case I plus…

- size challenges starts here - fully addressed in Pilot Phase
- API access - should be final by the end of Prototype Phase
  - simple cases should be fully automated by the end of phase
- get in touch with ‘tapes’!
- inheritance/dynamic handling of metadata schemas/definitions
  - communities need time to learn and find appropriate schema(s)
- initial local/hybrid deployments - k8s cluster, object store and tape exist
- segregation of config/planning and creation of archives
- simple ‘open data’ scenarios
Case III - large size - EuXFEL lab / extension of Case II

- Case II plus...

- petabytes range/millions of files here
- non interactive/human driven, except configuration
  - automated execution from day one - expect APIs ‘near ready’ by the end of prototype phase
- stacked/inherited data preservation policies (site -> lab -> experiment)
  - not strictly bound to metadata schema structure
Extras

- immutability of archives, but possibility to make changes efficiently (versions, deltas, ...)
- flexible hybrid deployment schemes (e.g. meta on-prem, data in the cloud or one copy on-prem, one copy in the cloud)

in one line - fruitful and productive months ahead!
EMBL-EBI Requirements and Expectations

Tony Wildish
Data volume doubles every two years

• => half of our data will always be < 2 years old
EBI data: *almost* Archival...

• Our data doesn’t go cold as fast as in other domains
  • Data volume doubles every two years,
    • So half our data is < 2 years old
  • A typical research project can last 2-5 years
    • …therefore…
  • Expect most of our data to be in active use, at some level, all the time

• Older data still has value
  • Tracking the rate of mutations in a virus
  • Tracking the spread of a gene through a population over time
  • Longitudinal studies, tracking people’s health throughout their life
Dataset definitions…

• A single research topic can use data from many other studies

• ~100K life scientists in Europe alone, all using similar data in different ways

• Dataset definitions overlapping, not orthogonal

• Highly dynamic!
EBI use-cases:

• Our use-cases are about managing the coldest data
  • Active -> cached -> archive, and back to active
  • How to identify the colder data -> biggest driver for costs
  • How to manage it cost-effectively?

• Our testing will focus on *dynamic* use of the archives
  • Ingestion rates (data, and metadata)
    • Data limited by h/w, metadata less so
  • Data migration between tiers, both up and down
    • Driven by user-activity, automated
  • Metadata operations: defining datasets, updating them
    • ~every research question will be a new dataset
PIC Requirements and Expectations

J. Casals, M. Delfino, J. Delgado
Port d’Informació Científica

Use cases will be based on MAGIC Telescope data at Observatorio del Roque de los Muchachos (La Palma, Canary Islands)

- Collecting data 365 days a year
- 300TB per year for ranges of 5-6 years
- Random recalls during the period

---

In collaboration with ALBA Synchrotron

- More than 10 Beamlines (and growing for next years)
- Datasets ranging from 200TB up to 4PB
- Internal and external scientific users
Prototype Requirements

- Petabyte level Storage → functional, reliable, good performance, reasonable cost
  - From 1PB in 2021 to 15PB in 2025
  - GEANT connection with bandwidths from 1Gbps - 10Gbps and up to 1Gbps in avg for 24h
  - Bulk download and upload → No price per file operation
- Actionable by automated data management scripts at datacenter → CLI and (at least) API
- Metadata driven Data Management and Data Archiving and Preservation
  - Possibility to create custom metadata schemas
- Integration with external identity providers → eduGain and Elixir at least
- Fine granularity permission control for data access and distribution
  - Offline and cloud embargo periods plus public distribution
- Ability to do in-archive data processing using co-located Cloud
  - Enables automatic processing of uploaded data
  - Prevents downloading processing and uploading it again
- Future reprocessing (reusable) possibilities (container/notebook systems so data doesn’t “expire”)
Early Adopters Programme

Anna Manou - CERN
Early adopters Programme

WHAT?

WHY?

HOW?
Public sector & not-for-profit organisations interested in the ARCHIVER PCP

Help to shape the R&D

Test the solutions developed

Potential to purchasing pilot-scale services
Becoming an Early Adopter means:

- Be consulted during the preparation of future ARCHIVER phases
- Access material produced by the project
- Propose your own use cases and get the chance to test resulting services
- Benefit from training sessions covering the services developed during the ARCHIVER project
- Accelerate the procurement process of pilot-scale services & have certain conditions
What are the obligations as an Early Adopter of ARCHIVER?

- Sign a **declaration of confidentiality** and **non-conflict of interest**, stating that your organisation will not submit a bid in response to the ARCHIVER Request for Tender.

- Allow the ARCHIVER Buyers Group to list your organisation's name in its **Request for Tenders** and subsequent **Call-offs**.

- In case of engagement in testing activities, **describe the use case(s)** to potentially test using the ARCHIVER services and to **provide structured feedback on the testing results** to the ARCHIVER project.

- Acknowledge the support of the European Commission and ARCHIVER project in any publications that result from the aforementioned testing activities performed with the developed services.
The Early Adopters engaged so far
Use cases

Archival and accessibility of omics data

Archiving Genomic and Imaging Data

Multi-Repository Research Data Harvester and Transformer for Swedish Archival Standard

Preserving Australia’s digital research, education and cultural heritage

Defining National Scale Data Archive Services

https://archiver-project.eu/early-adopters-use-cases
Are you part of a public sector research organisation with needs for standards-based, cost-effective data archiving and preservation services?

Are high ingest rates, data volumes at scale and long-term support important to you?

Express your interest

September 2021
Do you want to know more about the Early Adopters Programme?

https://archiver-project.eu/early-adopters-programme
BREAK

Video: Outcomes of the design phase
PROTOTYPE PHASE

AWARD CEREMONY
T-Systems International – GWDG – Onedata
Cloud-agnostic approach for research

https://tinyurl.com/y23gskwj
ARCHIVING AND PRESERVING TO DISCOVER

PROTOTYPE PHASE

Jurry de la Mar

T·Systems
“WE MOBILIZE MORE KNOW-HOW AND CREATE MORE DATA DISCOVERY IN RESEARCH BY DEMOCRATIZING ACCESS TO PROFESSIONAL ARCHIVING AND PRESERVATION FOR THE COST OF STORING THE INFORMATION.”

Team T-Systems.
INNOVATE, PERFORM AND SUSTAIN
T-Systems Team of Experts

Jurry de la Mar
Science and Research Expert, T-Systems

Lukasz Dutka
Research Data Expert, Onedata

Prof. Dr. Philipp Wieder
Research Data and Preservation Expert, GWDG

Dietrich Singh
Cloud Expert, T-Systems

Bartosz Kryza
Distributed Data Expert, Onedata

Prof. Dr. Ramin Yahyapour
Research Data and IT Expert, GWDG
Modern data access and management platform allowing automated push/pull ingestion from legacy data sources

Creation of arbitrary size standards compliant archive-packages in cost- and energy-effective manner

Highly scalable and flexible metadata extraction framework

Flexible setup and control of archiving pipelines with BPMN workflows

Fully open-source and deployable on Kubernetes

BPMN = Business Process Model and Notation Standard
# Prototype Focus

<table>
<thead>
<tr>
<th>Functionality</th>
<th>Prototype Phase</th>
<th>Pilot Phase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sustainability</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Hybrid Deployments</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Data redundancy</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Data-at-rest encryption</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Data immutability</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Extended KPI monitoring</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>CTS/ISO certification</td>
<td>Preparation</td>
<td>X</td>
</tr>
<tr>
<td>Intelligent caching</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Cloud UI integration</td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>

The proof of the pudding is in the eating!
COMMERCIALISATION AND SUSTAINABILITY

- Freemium service
  - Support EOSC
  - Long-tail of science
  - Quick uptake
- Premium service
  - Advanced and customized solutions
- Collaborate model
  - On-premise services
  - 3rd party service providers
  - Collaboration with GAIA-X
  - Continuous innovation and sustainability of toolsets

The Deutsche Telekom Group-wide program "We care for our planet" objective is to help the company achieve its climate targets: 100% green energy by 2021, 90% emission reductions by 2030, carbon neutral by 2050.
THANK YOU

LIFE IS FOR SHARING.
Building the next generation Research Data Management solution
• LIBNOVA mission is to safeguard the world’s research and cultural heritage. Forever.

• LIBNOVA is a world leader in digital preservation, was founded in 2009, has offices in Europe and the US and is now present in 14 countries with activity in the academic, cultural heritage and research communities.

• Customers like the British Library, HILA Stanford University, the EPFL and many more already trust us.
• The University of Barcelona is the foremost public institution of higher education in Catalonia, catering to the needs of the greatest number of students and delivering the broadest and most comprehensive offering in higher educational courses.

• The University of Barcelona is also the principal center of university research in Spain and has become a European benchmark for research activity, both in terms of the number of research programmes it conducts and the excellence these have achieved.
The Spanish National Research Council is the main agent of the Spanish System for Science, Technology and Innovation with competences aimed at: Generation of knowledge through scientific and technical research, Transfer of results from research, especially to boost and create technology-based enterprises, Expert advice provided to public and private institutions, Highly-qualified pre-doctoral and post-doctoral training, Promotion of scientific culture in society and management of large facilities and unique scientific and technical infrastructures.
David Giaretta has worked in digital preservation since 1990 and has led many of the most important developments in this area. He chaired the panel which produced the OAIS Reference Model (ISO 14721), the “de facto” standard for building digital archives, and made fundamental contributions to that standard. He leads the group which produced the ISO standard for audit and certification of trustworthy digital repositories (ISO 16363), and ISO 16919.

Involved with the Alliance for Permanent Access (APA) from its start to its establishment, he became the Director of the APA in July 2010.
Amazon Web Services (AWS) is the world’s most comprehensive and broadly adopted cloud platform, offering over 175 fully featured services from data centers globally.

Millions of customers are using AWS to lower costs, become more agile, and innovate faster. AWS has the most extensive global cloud infrastructure. With multiple Availability Zones connected by low latency, high throughput, and highly redundant networking. AWS has 77 Availability Zones within 24 geographic regions around the world.
Voxility provides agile Infrastructure-as-a-Service in the biggest Internet hubs in the world: when, how and where is needed.

Massive scalability, raw processing power and the faster network connections across the world.
• Multi-petabyte scale with the **CSIC**’s vast experience on supercomputing and large-scale infrastructures plus **Amazon Web Services** and **Voxility Infrastructure**.

• Fully aligned with the EU legal requirements, GDPR, FAIR principles, TRUST principles and applying really advanced Artificial Intelligence techniques to gain unprecedented efficiency (classification, PII detection, etc) working with the **Universitat de Barcelona**.

• Completely aligned to the OAIS, ISO 16363 and CoreTrustSeal, working with **David Giarietta**.

• Built on top of **LIBNOVA**’s rock-solid foundation, based on our extensive digital preservation experience and proven solutions, already running in the most demanding organizations worldwide.
LABDRIVE: Research data management

Research organizations need to:

- Be confident about how research data is managed and protected for the whole data lifecycle, capturing it as soon as possible.

- Provide the best available tools for their researchers, carefully balancing resources across research projects.

LABDRIVE is the foreseen solution with which organizations will create the research data they produce and keep it protected, for all their projects/units/departments, starting when the data is created and for the long term, in a single platform.
LABDRIVE will be:

- **Long-term preservation oriented:** OAIS, ISO16363, TRUST, FAIR, PREMIS among others are at the core of the solution.
- **Performant:** Scalable and parallelized, capable of preserving in the petabyte scale.
- **Flexible:** As a service or on-premises
- **Multi cloud:** No cloud vendor lock-in, open to competition.
- **Multi-protocol:** S3, rsync, SFTP, NFS and other protocols can be used to access data.
- **Interoperable:** Extensive API plus the adoption of open standards: BagIt, METS, Premis, etc.
- **Environmentally friendly:** Runs with minimal environment impact.
Approach

- **Multidisciplinary:** Our consortia includes the University world, public research center, best field experts, infrastructure leaders and a world leader in preservation.

- **Disruptive:** We are thinking long-term. How to change the approach to solve the challenge.

- **Co-Developed:** Working together with the Buyers Group and early adopters, to understand their needs and create best practices.
Contact LIBNOVA:
contact@libnova.com

Contact me:
a.guillermo@libnova.com
Arkivum - Google
Google Cloud Platform for research

https://tinyurl.com/y4ycaqpa
ARCHIVER
Arkivum and Google solution
Phase 2: Prototype
Arkivum / Google Solution:

- Scalable storage and compute
- High speed ingest and access
- Policy based cost optimization
- OAIS workflows and packages
- Digital Preservation rules and actions
- FAIR datasets and access
- Hosted scientific applications
- Open standards and specifications
- Exit and migration strategies
Google Cloud Platform: PB Scale Storage, Compute and Networking

- Google Object Storage
- Google File Storage
- Google Compute Engine
- Google Kubernetes Engine
- Google Operations
- Google Security
- High speed network
- GEANT connected
Prototype: Portability and Exit Strategies

- Deployment in GCP, on-premise and hybrid cloud
- Portable to other cloud providers
- Kubernetes, containers, Anthos, automated deployment
- Exit strategies using data escrow, open standards and fast exports
Pilot: Long Term Digital Preservation Hosted On GCP
Prototype: Factories for LTDP in Large Scale Science

Content types and sources

Automated Workflows

- CERN CMS Open Data Workstream
- PIC Telescope Workstream
- EBI Genomics Workstream
- DESY Synchrotron Workstream
- CERN Digital Memory Workstream

FAIR data for Researchers

API

images: Flaticon.com
Prototype: Approach

- Automation, Scalability and Efficiency: Preservation Factories
- Minimal Effort Ingest / Minimal Viable Preservation
- Dataset Authenticity, Integrity and Usability: FAIR
- Platform for building Trusted Digital Repositories
- Fully SaaS on GCP, but also portable to on-premise and hybrid deployments
Thank you

https://www.archiver-project.eu/

www.arkivum.com

Find us on LinkedIn or on Twitter @Arkivum
Feedback Session

Marion Devouassoux - CERN
Questions

1. What is your role in this award ceremony?
2. This award ceremony helped me better understand the project. Do you agree?
3. Did you receive sufficient information on the selected consortia's planned solutions?
4. Do you expect the ARCHIVER resulting services to meet your needs?
5. Is the EOSC-Exchange a good channel to make available the resulting ARCHIVER services to the wider research community?
Go to menti.com

- Grab your phone or open a new window
- Go to www.menti.com