ARCHIVER

ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

PROTOTYPE PHASE KICK-OFF EVENT AND AWARD CEREMONY

07 December 2020

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

ARCHIVER - Archiving and Preservation for Research Environments project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824516.

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 Devouassoux (CERN)

ARCHIVER

ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

European Molecular Biology Laboratory
European Bioinformatics Institute

The home for big data in biology

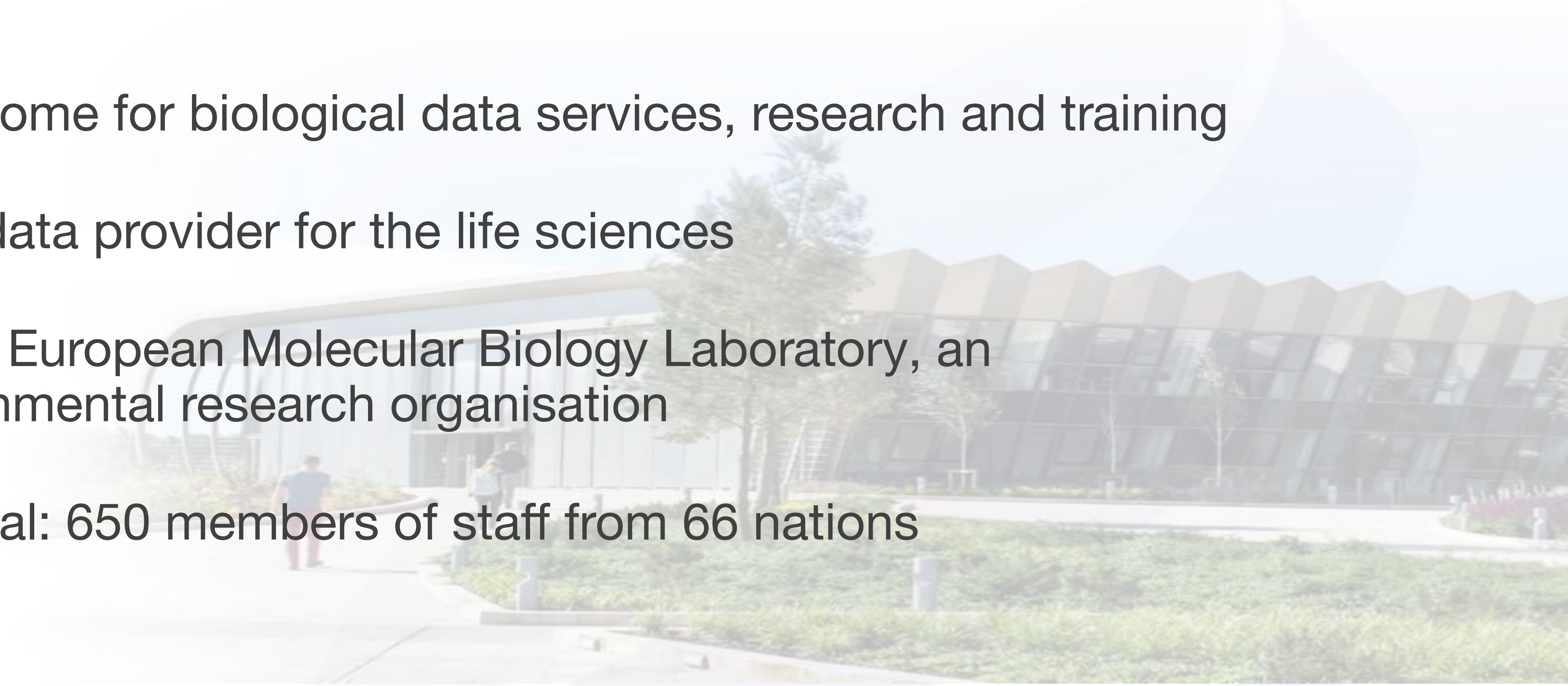
Welcome!

Prototype Phase Public Awards Ceremony
December 7th 2020

Tony Wildish – EMBL-EBI

ARCHIVER - Archiving and Preservation for Research Environments project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824516.

What is EMBL-EBI?



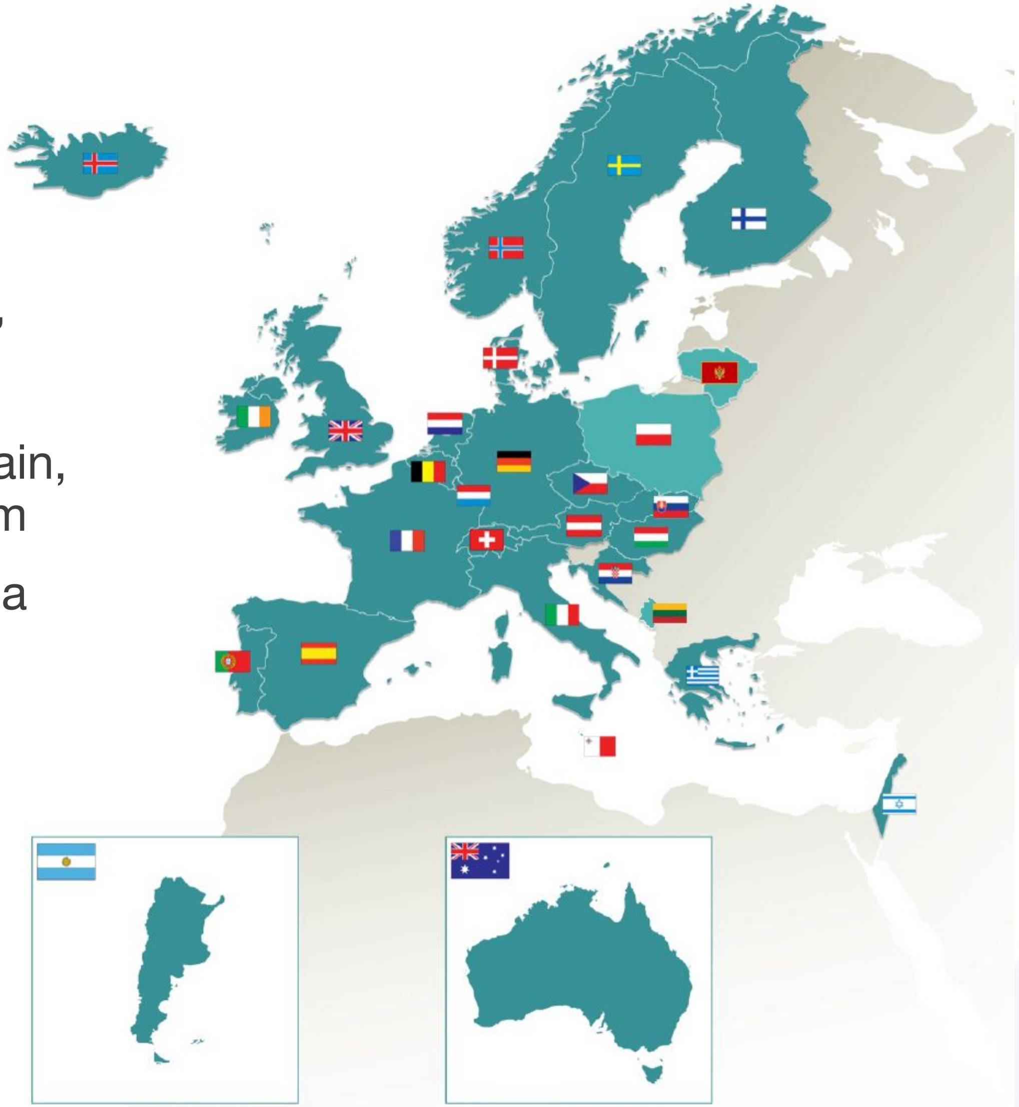
- Europe’s home for biological data services, research and training
- A trusted data provider for the life sciences
- Part of the European Molecular Biology Laboratory, an intergovernmental research organisation
- International: 650 members of staff from 66 nations

EMBL member states

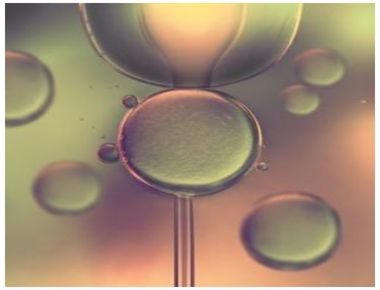
Austria, Belgium, Croatia, Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Luxembourg, Malta, Montenegro, the Netherlands, Norway, Portugal, Slovakia, Spain, Sweden, Switzerland and the United Kingdom

Associate member states: Argentina, Australia

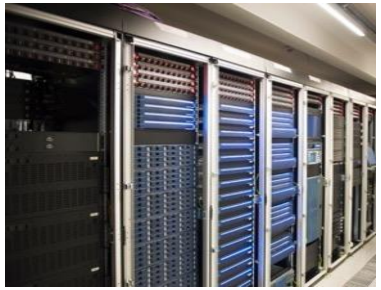
Prospect member states: Lithuania, Poland



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

>1700 personnel

6 sites in Europe

Heidelberg, Germany

Main Laboratory

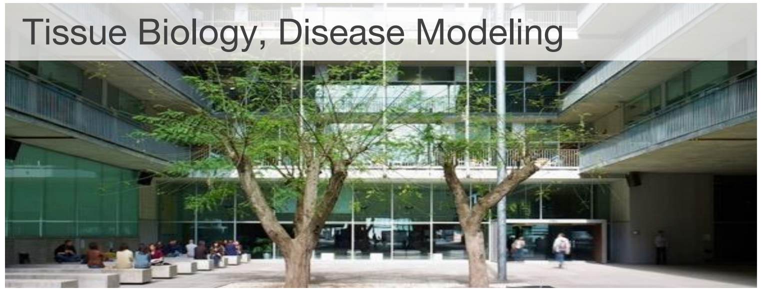
Hinxton, Cambridge, UK

Bioinformatics

Grenoble, France

Structural Biology

Tissue Biology, Disease Modeling



Barcelona, Spain

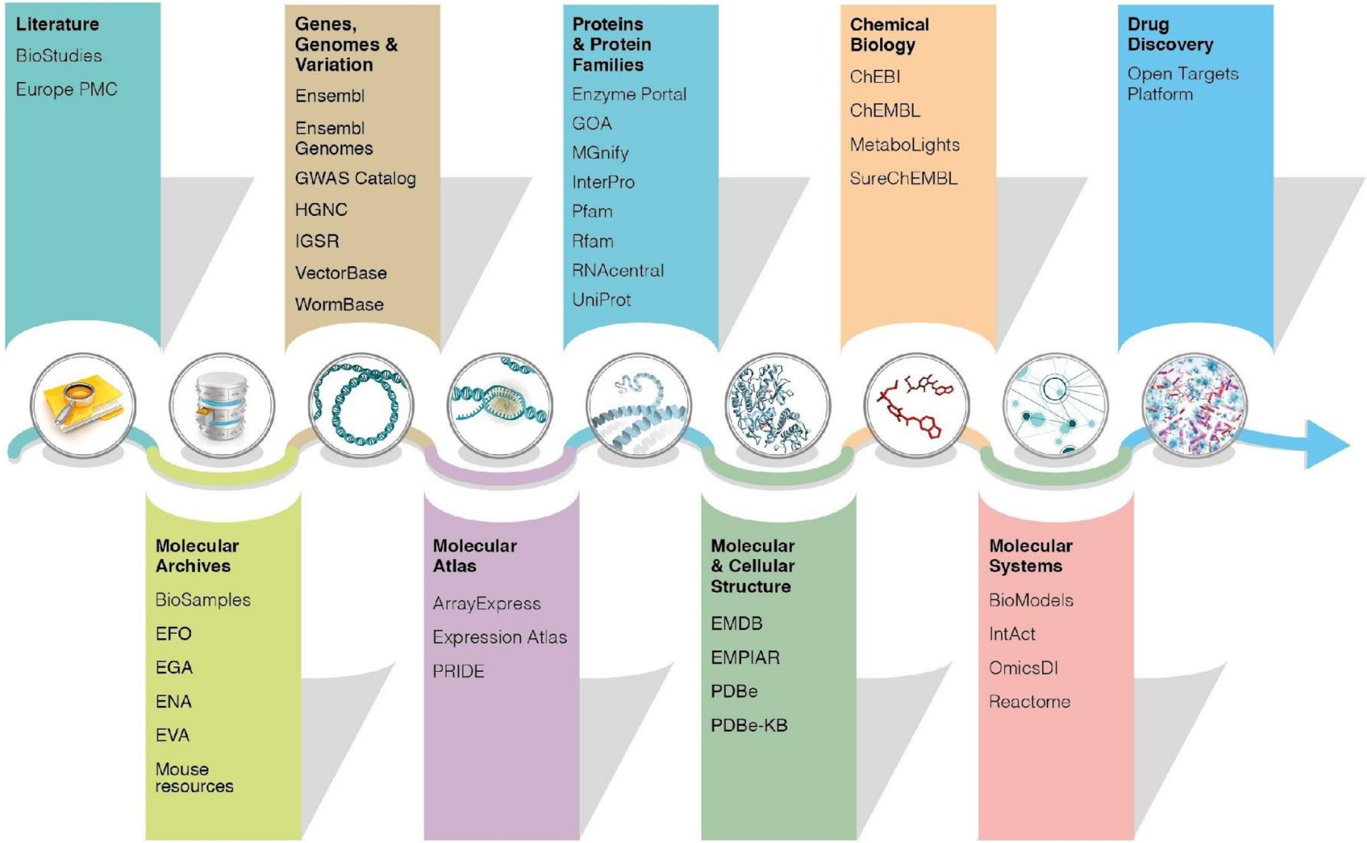
Mouse Biology

Rome, Italy

Structural Biology

Hamburg, Germany

Data resources at EMBL-EBI



Literature
BioStudies
Europe PMC

Molecular Archives
BioSamples
EFO
EGA
ENA
EVA
Mouse resources

Genes, Genomes & Variation
Ensembl
Ensembl Genomes
GWAS Catalog
HGNC
IGSR
VectorBase
WormBase

Molecular Atlas
ArrayExpress
Expression Atlas
PRIDE

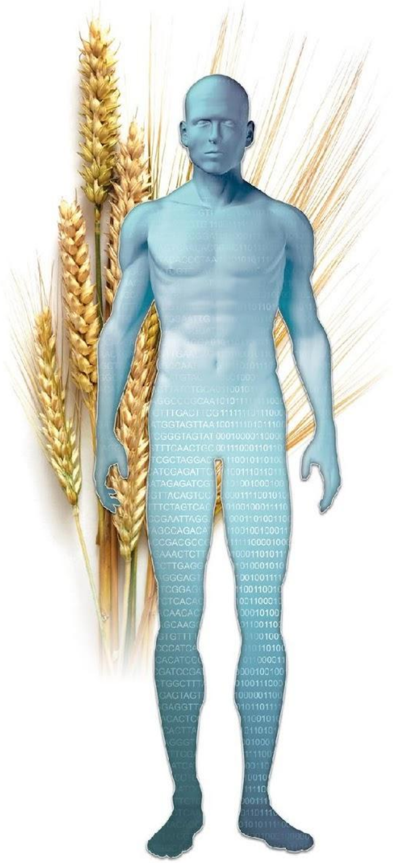
Proteins & Protein Families
Enzyme Portal
GOA
MGnify
InterPro
Pfam
Rfam
RNAcentral
UniProt

Molecular & Cellular Structure
EMDB
EMPIAR
PDBe
PDBe-KB

Chemical Biology
ChEBI
ChEMBL
MetaboLights
SureChEMBL

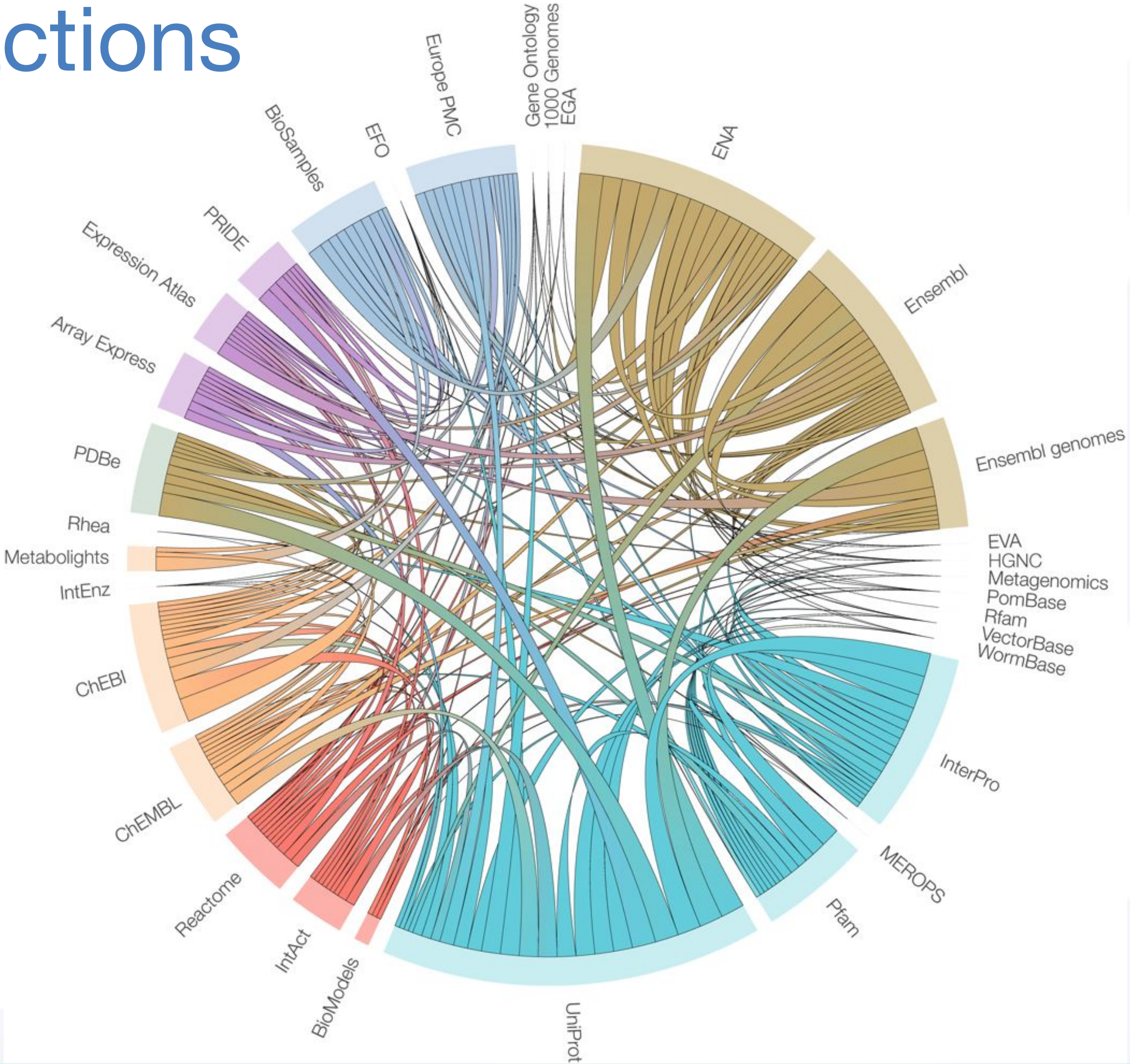
Molecular Systems
BioModels
IntAct
OmicsDI
Reactome

Drug Discovery
Open Targets Platform

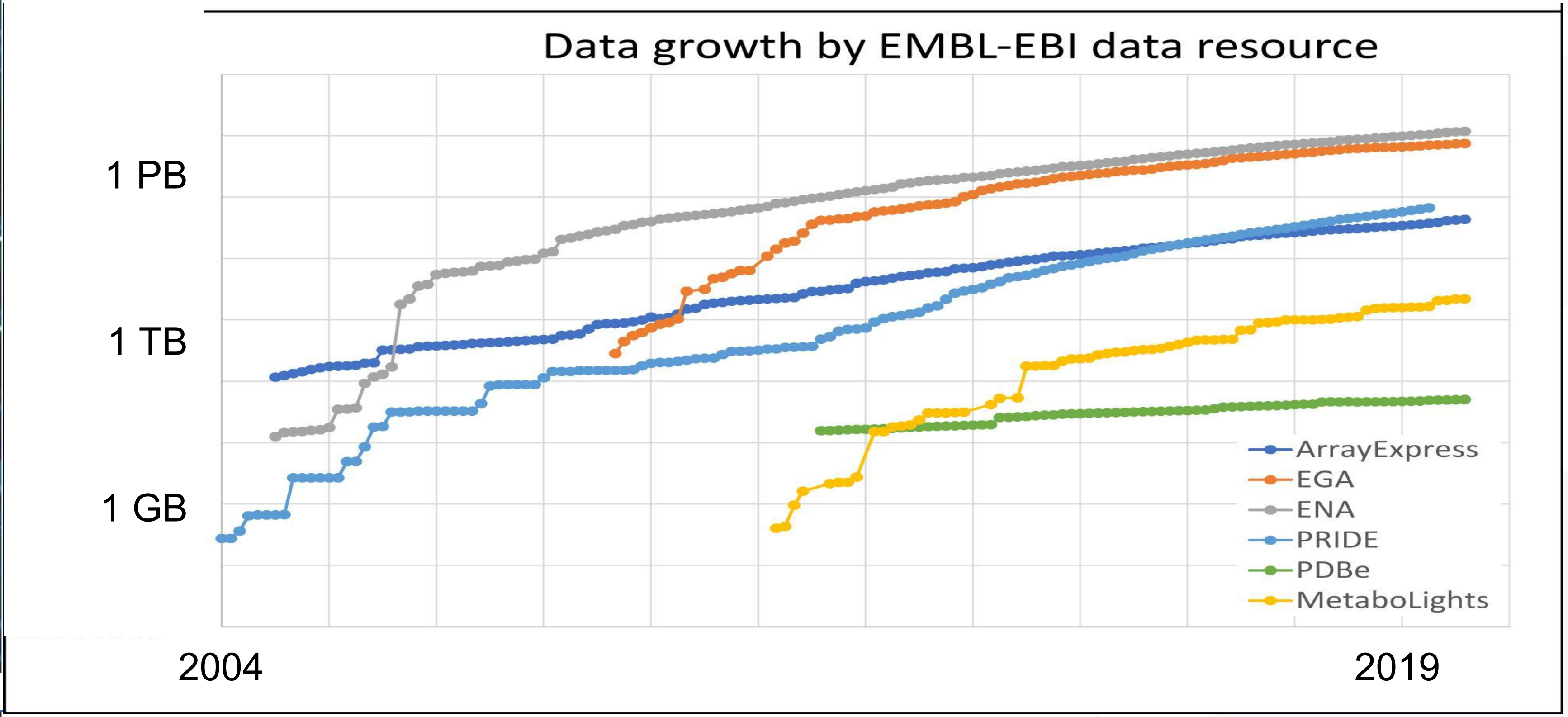


Database interactions

- Data exchange between EBI data resources
- Arc width weighted by the number of different data types exchanged



Increasing Data, Increasing Analysis



Data volume doubles every two years

- => half of our data will always be < 2 years old

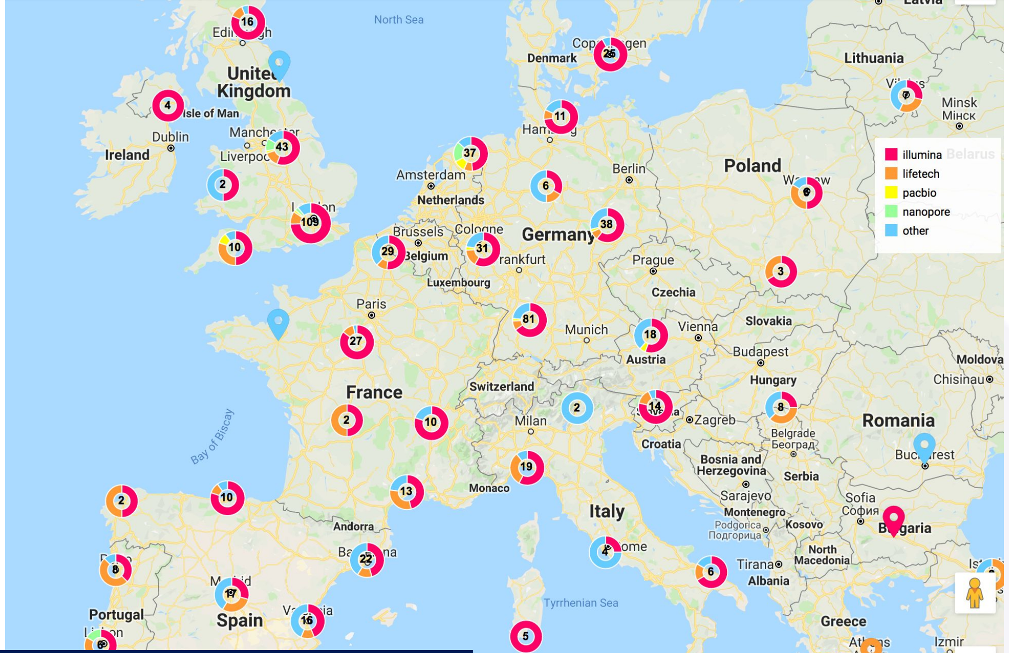
EGA and ENA account for the bulk of the data

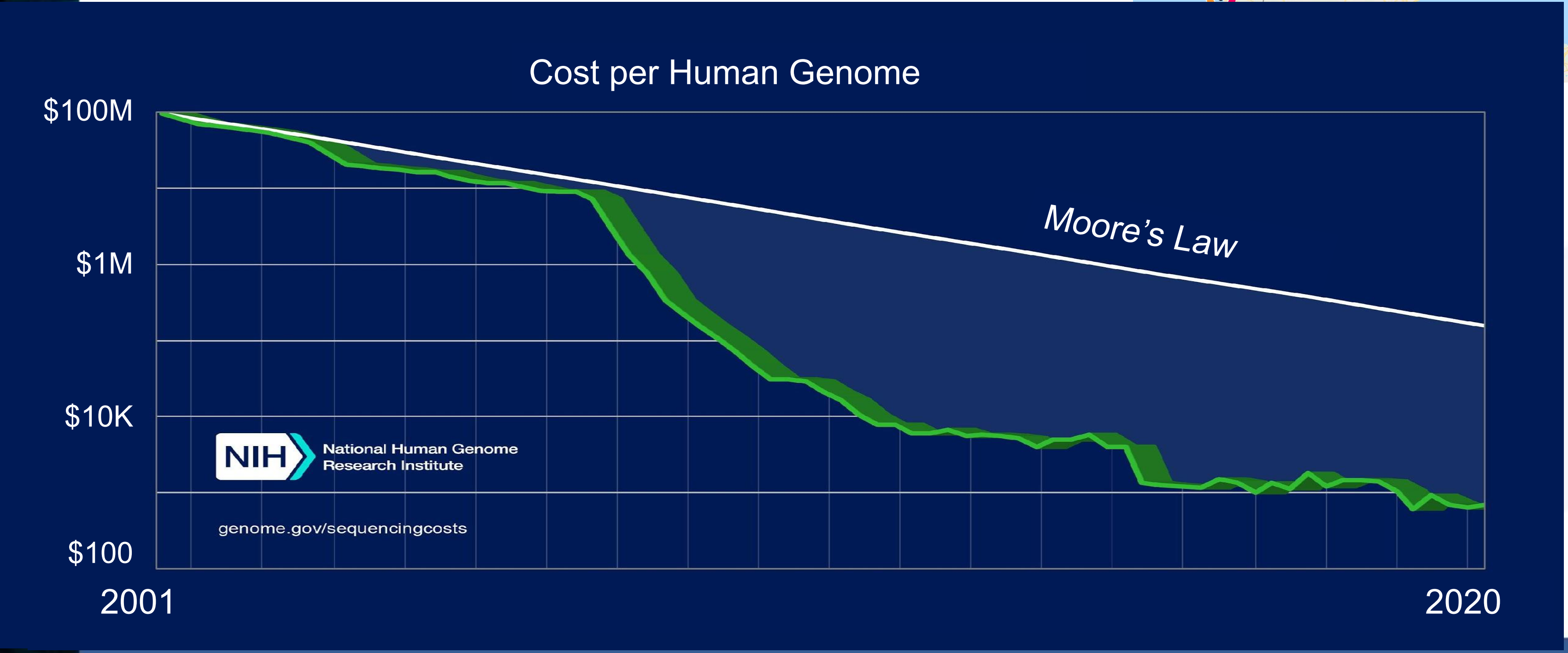
- DNA sequences

BioImaging repository

- Just starting, will be big

Our data comes from everywhere

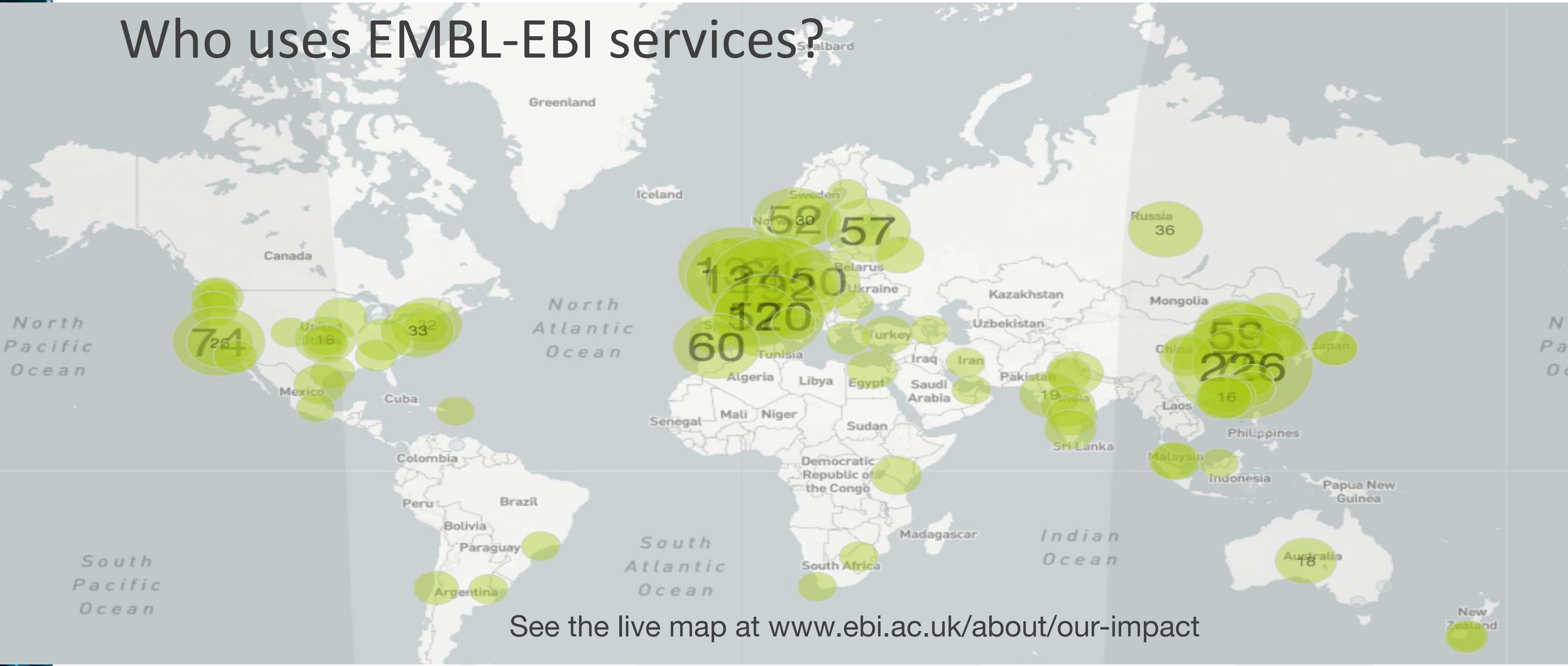




genome.gov/sequencingcosts

And is getting cheaper to produce

Who uses EMBL-EBI services?



See the live map at www.ebi.ac.uk/about/our-impact

EMBL-EBI

EMBL-EBI
EMBL-EBI
Celebrating
20 years of
bioinformati
MBL-EBI
OUTH
UILDING

Video: Outcomes of the design phase

ARCHIVER
Pre-Commercial Procurement
Multiple ICT technologies in a hybrid cloud environment
0:19 / 2:16

ARCHIVER

HELIX NEBULA

ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

Project overview / update

João Fernandes – CERN

ARCHIVER - Archiving and Preservation for Research Environments project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824516.

Project

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

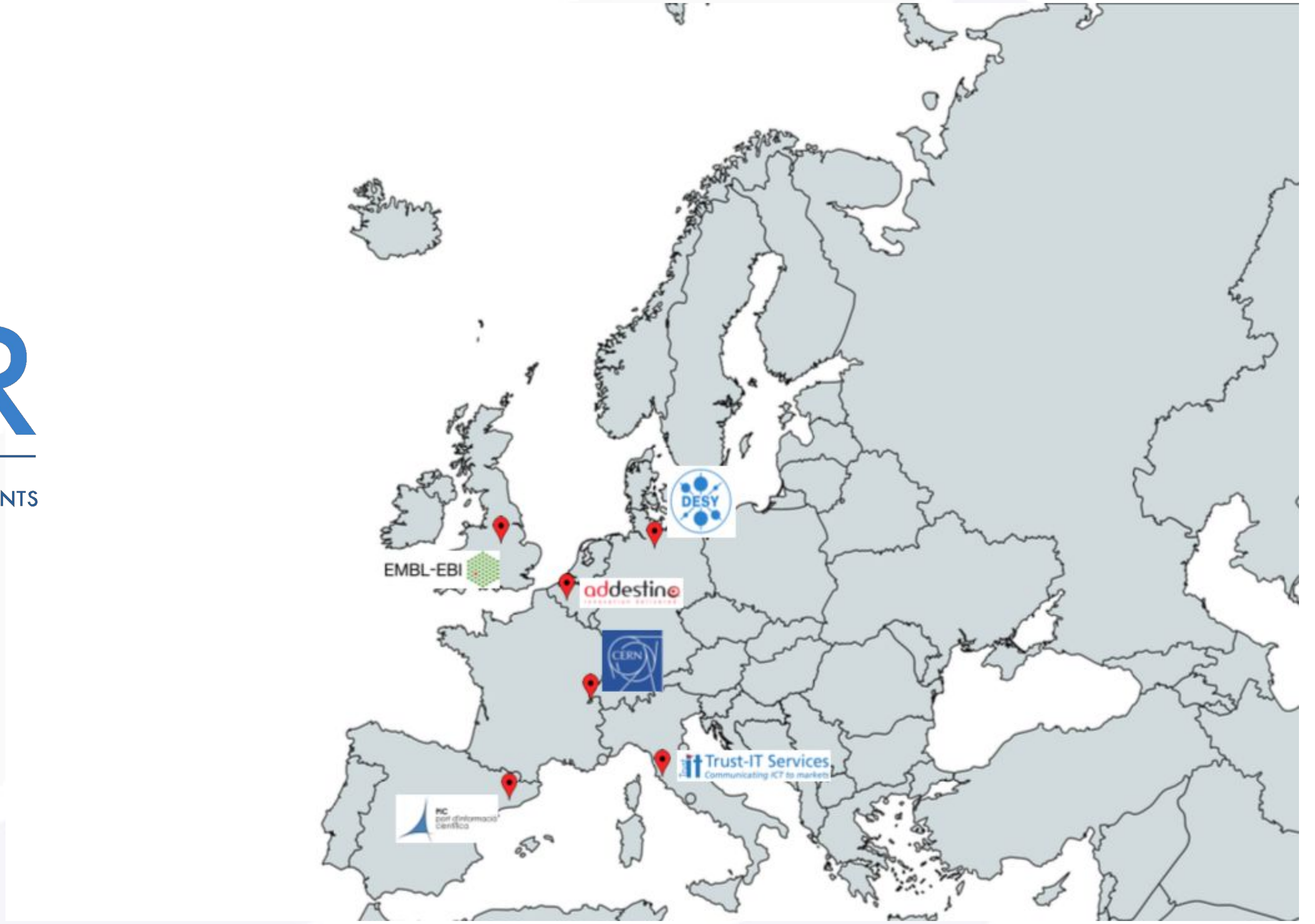
Starting Date: 1st of January 2019

Duration: 42 Months

Coordinator: CERN (Lead Procurer)

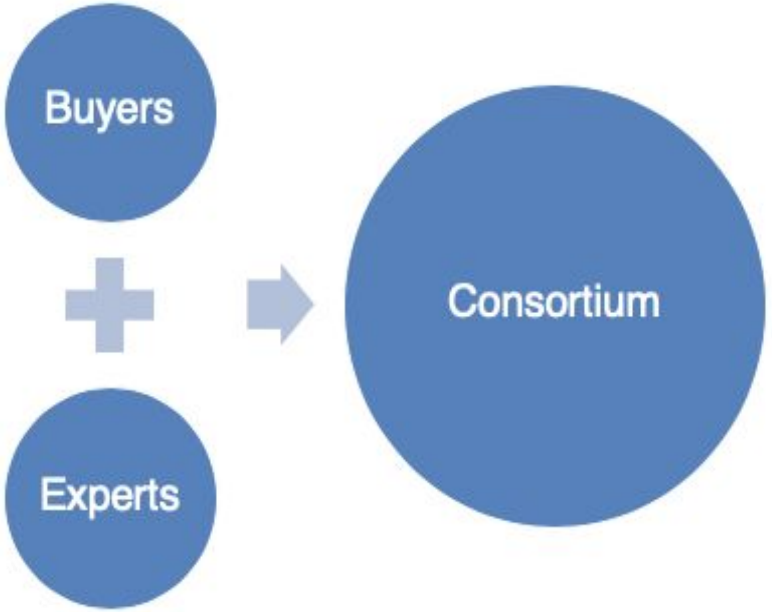
ARCHIVER
ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

European Commission



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

Current Scientific Data Repositories

ARCHIVER
ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

Growing data volumes	→	*PB scale demonstration of scientific data repositories*
Basic bit preservation capabilities	→	*European SaaS providers in digital preservation*
Most of research data not available	→	*Best practices: FAIR, TRUST, DPC RAM*
Technology lock-ins concerns (tape), Business Continuity plans needed (COVID-19)	→	*Promote FOSS, open standards & non bespoke services, demonstration of exit strategies*
Fragmentation across scientific disciplines & countries	→	*Pan-European: resulting services available in the EOSC portfolio*
Cost underestimation at the planning phase	→	*Cost model adapted to public research*

ARCHIVER "current state of the art" report: https://doi.org/10.5281/zenodo.3618215

R&D Scope

Demand Side Requirements

Layer 4
Advanced services

High level services: visual representation of data (domain specific), reproducibility of scientific analyses, etc.

Layer 3
Baseline user services

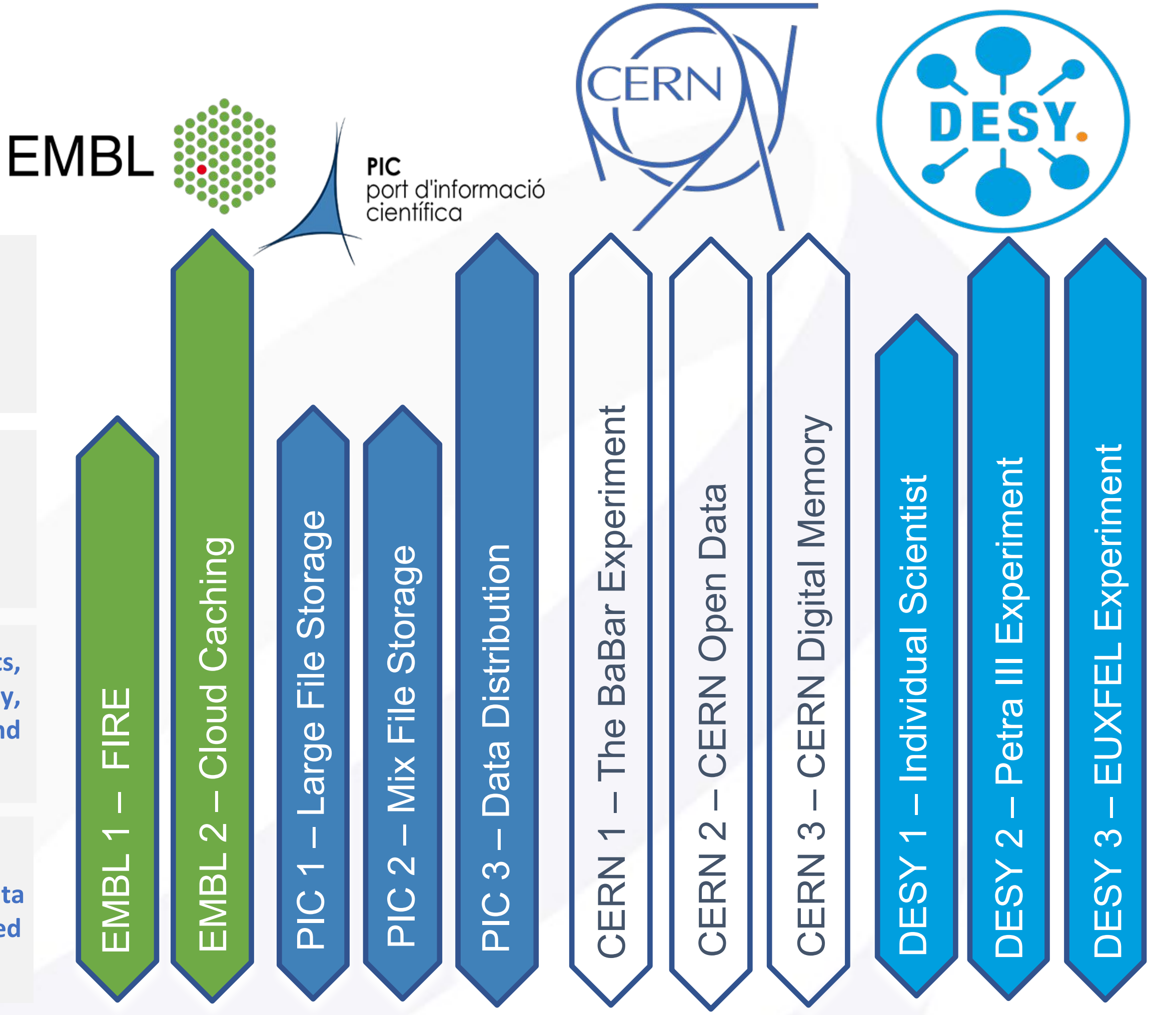
User services: search, discover, share, indexing, data removal, etc. Access under Federated IAM

Layer 2
Preservation

OAIS conformant services: data readability formats, normalization, obsolesce monitoring, files fixity, authenticity checks, etc. ISO 14721/16363, 26324 and related standards

Layer 1
Storage/Basic Archiving/Secure backup

Data integrity/security; cloud/hybrid deployment
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



EMBL 1 – FIRE

EMBL 2 – Cloud Caching

PIC 1 – Large File Storage

PIC 2 – Mix File Storage

PIC 3 – Data Distribution

CERN 1 – The BaBar Experiment

CERN 2 – CERN Open Data

CERN 3 – CERN Digital Memory

DESY 1 – Individual Scientist

DESY 2 – Petra III Experiment

DESY 3 – EUXFEL Experiment

Scientific use cases deployments documented at: 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

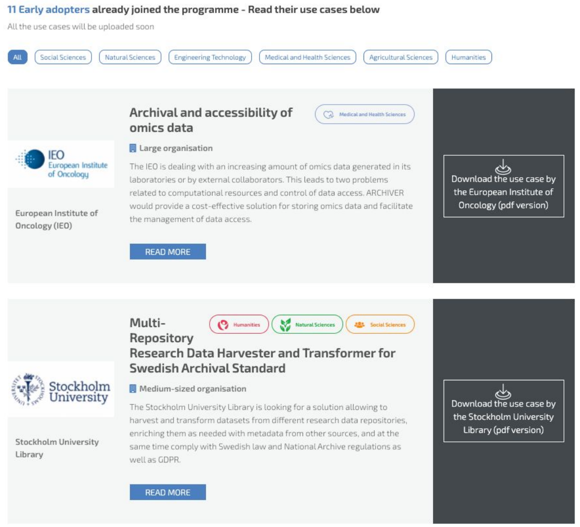
panosc
photon and neutron open science cloud

SSHOC
social sciences & humanities open cloud

LIFE IS NOT FAIR
OUR DATA IS
ENVRI
FAIR
WWW.ENVRI.EU

Early Adopters

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



11 Early adopters already joined the programme - Read their use cases below

All the use cases will be uploaded soon

Archival and accessibility of omics data

European Institute of Oncology (IEO)

Large organisation

The IEO is dealing with an increasing amount of omics data generated in its laboratories or by external collaborators. This leads to two problems related to computational resources and control of data access. ARCHIVER would provide a cost-effective solution for storing omics data and facilitate the management of data access.

Download the use case by the European Institute of Oncology (pdf version)

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

Stockholm University Library

Medium-sized organisation

The Stockholm University Library is looking for a solution allowing to harvest and transform datasets from different research data repositories, enriching them as needed with metadata from other sources, and at the same time comply with Swedish law and National Archive regulations as well as GDPR.

Download the use case by the Stockholm University Library (pdf version)

- **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**

CRG Centre for Genomic Regulation

IEO Istituto Europeo di Oncologia

SURF SARA

aarnet Australia's Academic and Research Network

SWITCH Serving Swiss Universities

KAUST

Universitet Stockholms

UKRI Science and Technology Facilities Council

ARDC Australian Research Data Commons

Jisc

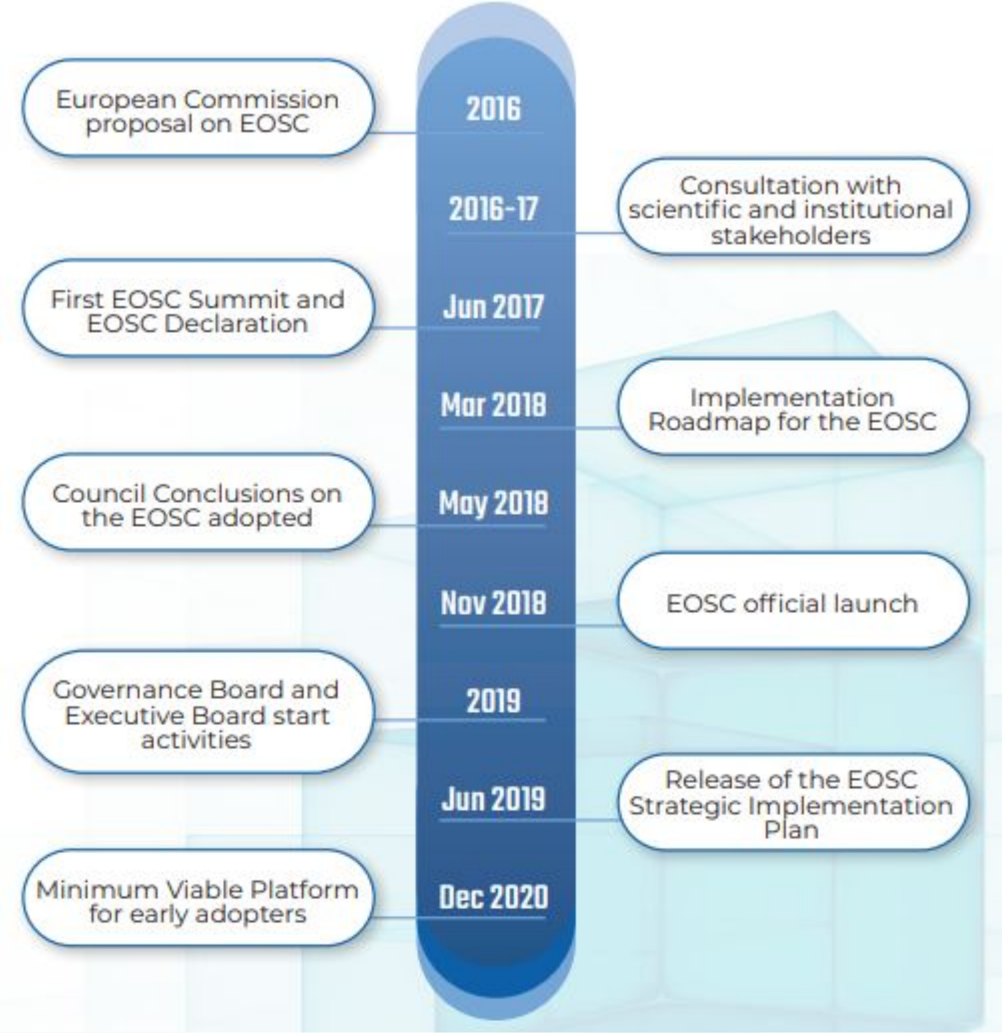
FMI Friedrich Miescher Institute for Biomedical Research

ALBA

European Open Science Cloud (EOSC)

> *"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



2016 – European Commission proposal on EOSC

2016-17 – Consultation with scientific and institutional stakeholders

Jun 2017 – First EOSC Summit and EOSC Declaration

Mar 2018 – Implementation Roadmap for the EOSC

May 2018 – Council Conclusions on the EOSC adopted

Nov 2018 – EOSC official launch

2019 – Governance Board and Executive Board start activities

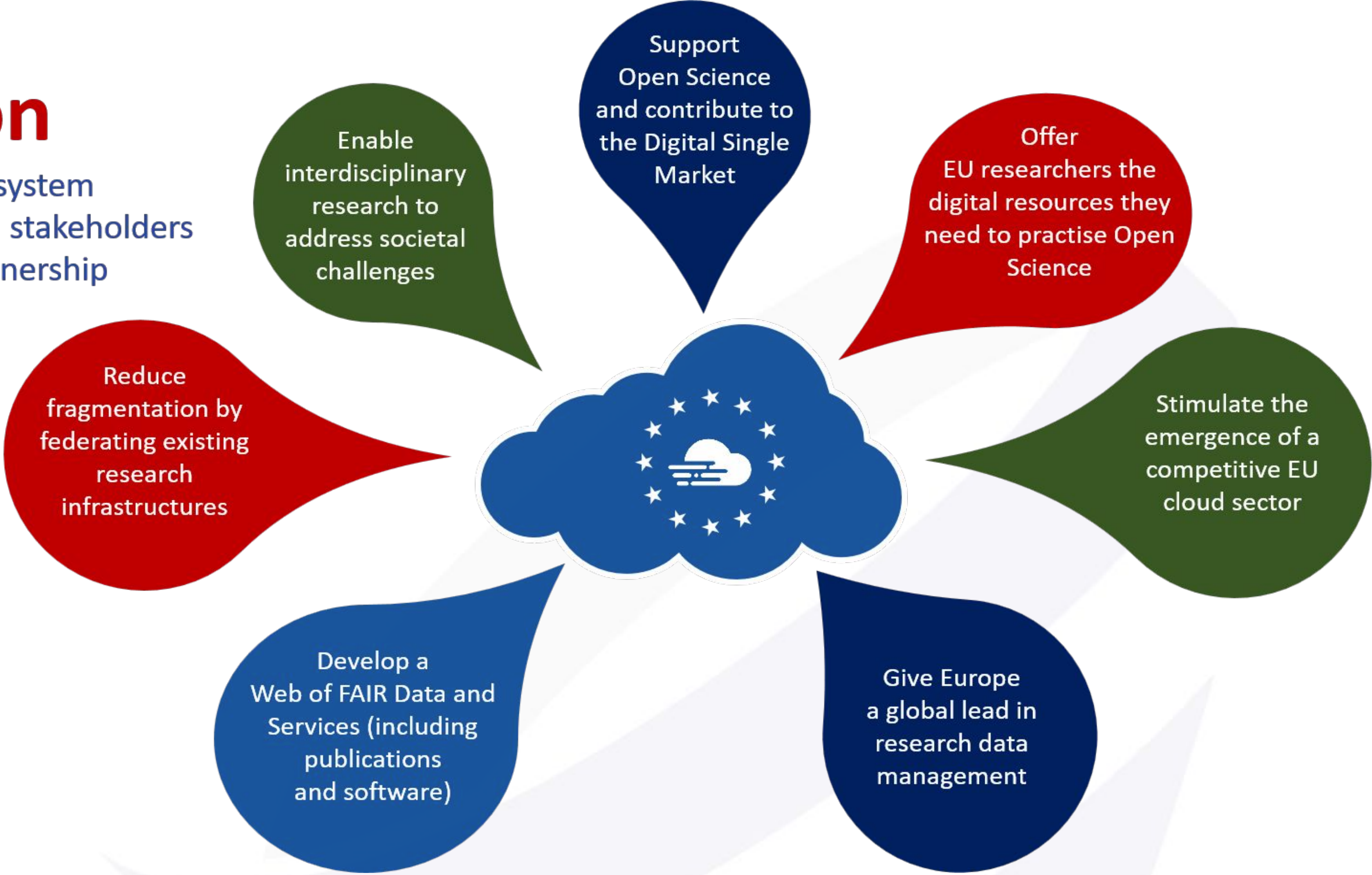
Jun 2019 – Release of the EOSC Strategic Implementation Plan

Dec 2020 – Minimum Viable Platform for early adopters

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

The Vision

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



Reduce fragmentation by federating existing research infrastructures

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

Stimulate the emergence of a competitive EU cloud sector

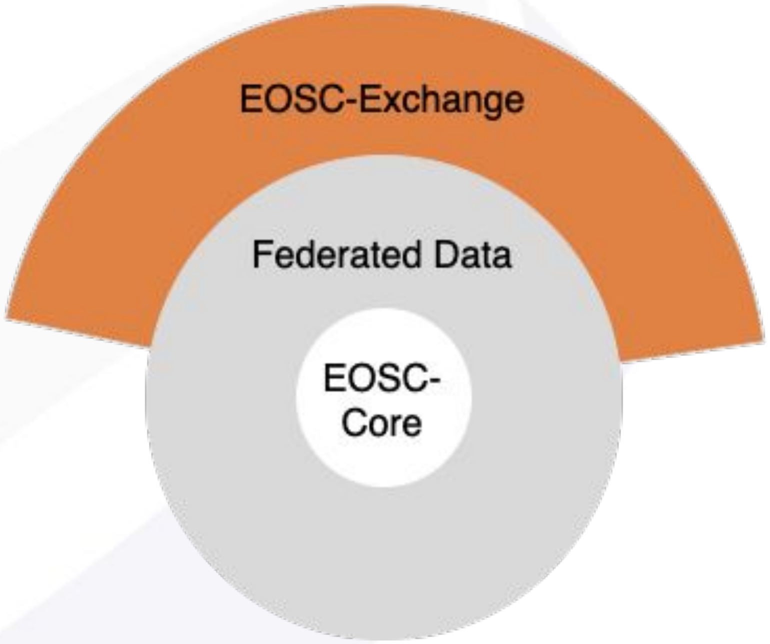
Give Europe a global lead in research data management

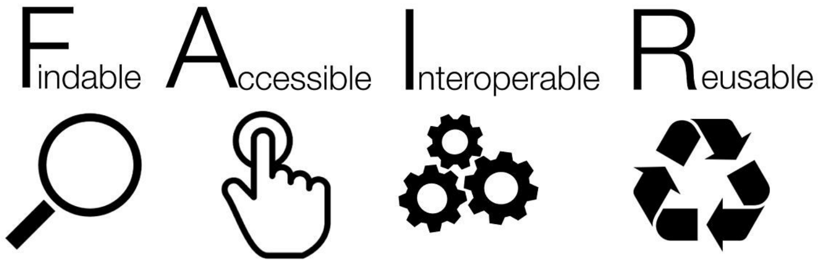
Develop a Web of FAIR Data and Services (including publications and software)

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





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

ARCHIVER & EOSC

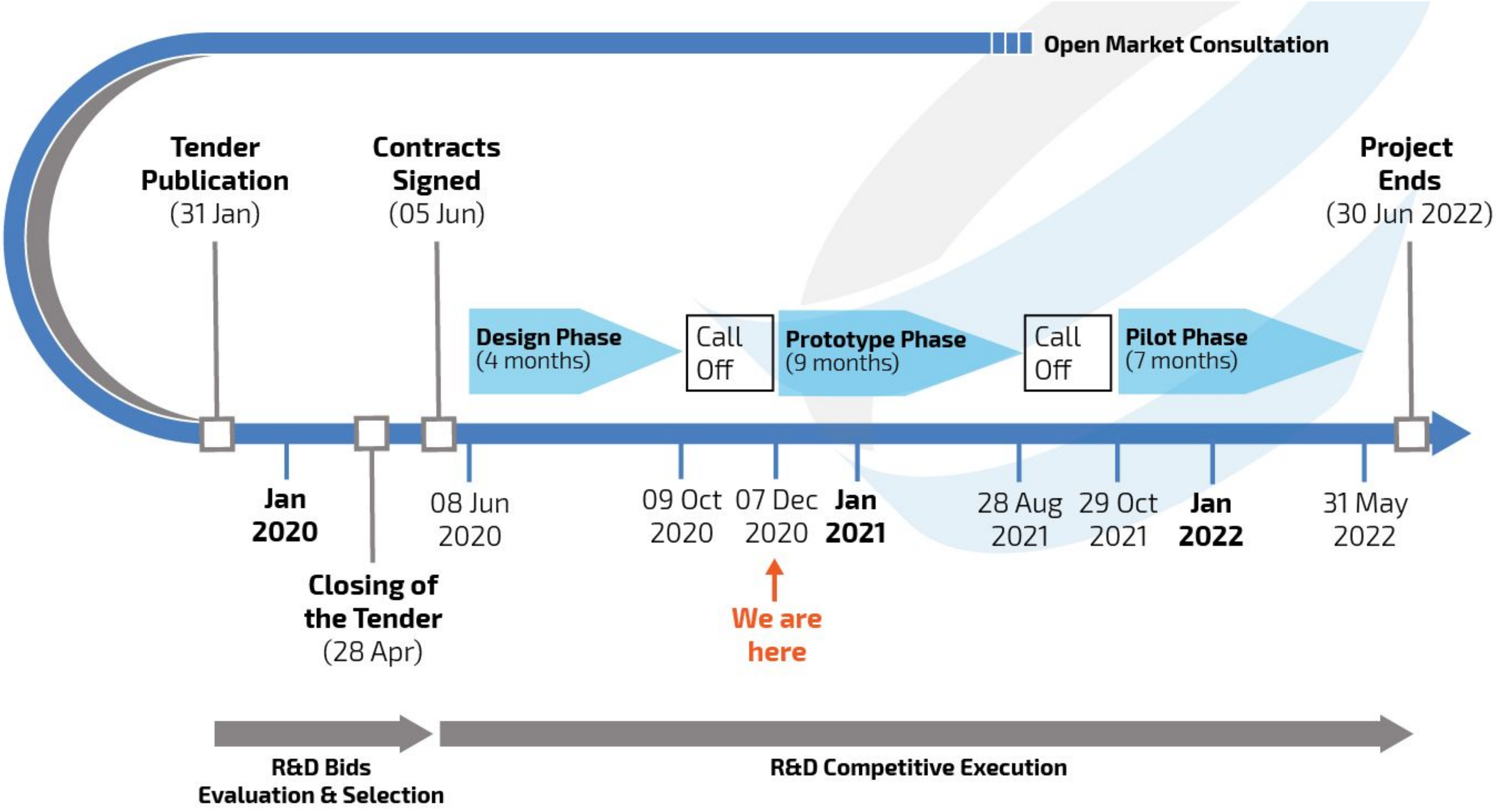
- Broad pan-European requirement analysis of the research sector
  - Analysis results considered in the competitive R&D tender
  - Technical and organisational measures aligned with European legislation in the services being developed (by default & by design)

- 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.

Timeline


Open Market Consultation
Tender Publication (31 Jan)
Contracts Signed (05 Jun)
Project Ends (30 Jun 2022)
Design Phase (4 months)
Call Off
Prototype Phase (9 months)
Call Off
Pilot Phase (7 months)
Jan 2020
Closing of the Tender (28 Apr)
08 Jun 2020
09 Oct 2020
07 Dec 2020
Jan 2021
We are here
28 Aug 2021
29 Oct 2021
Jan 2022
31 May 2022
R&D Bids Evaluation & Selection
R&D Competitive Execution

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

ONEDATA

GWDG
Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen

libnova

UNIVERSITAT DE BARCELONA

CSIC
CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

Giaretta Associates

amazon web services

voxility

arkivum
Bringing archived data to life

Google Cloud

https://www.archiver-project.eu/
https://twitter.com/ArchiverProject
https://www.linkedin.com/company/archiver-project/
https://www.youtube.com/channel/UCCBIyLpUt-hWmQatqdIhIzw

ARCHIVER

ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

Expected outcomes of the Prototype Phase

Buyers Group representatives

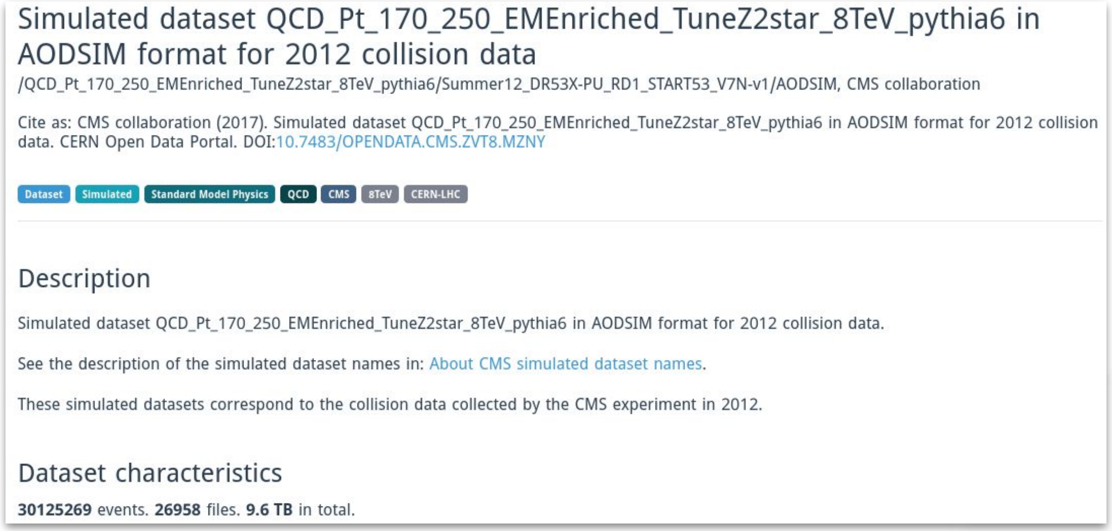
CERN, DESY, EMBL-EBI, PIC

ARCHIVER - Archiving and Preservation for Research Environments project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824516.

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_DR53X-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.ZVT8.MZNY

Description

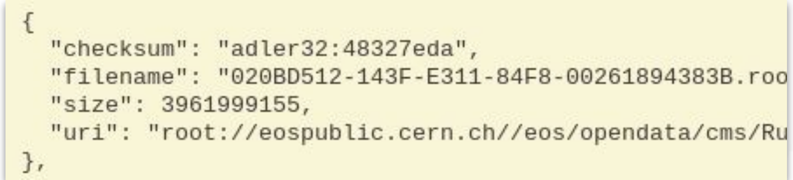
Simulated dataset QCD_Pt_170_250_EMEnriched_TuneZ2star_8TeV_pythia6 in AODSIM format for 2012 collision data.

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

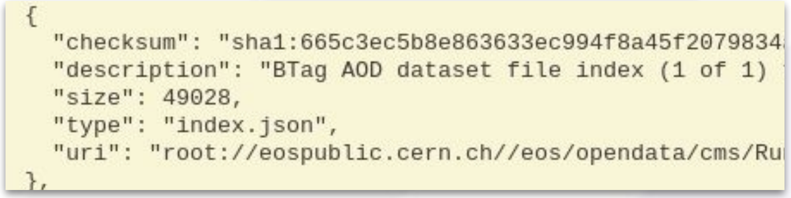
These simulated datasets correspond to the collision data collected by the CMS experiment in 2012.

Dataset characteristics

30125269 events. **26958** files. **9.6 TB** in total.

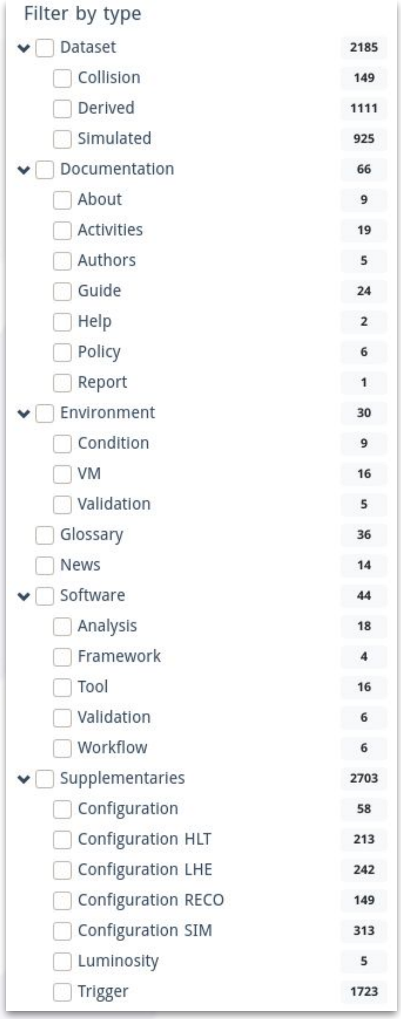


```
{
  "checksum": "adler32:48327eda",
  "filename": "020BD512-143F-E311-84F8-00261894383B.roo
  "size": 3961999155,
  "uri": "root://eospublic.cern.ch//eos/opendata/cms/Ru
},
```



```
{
  "checksum": "sha1:665c3ec5b8e863633ec994f8a45f2079834
  "description": "BTag AOD dataset file index (1 of 1)
  "size": 49028,
  "type": "index.json",
  "uri": "root://eospublic.cern.ch//eos/opendata/cms/Ru
},
```

Multiple checksumming options



Filter by type

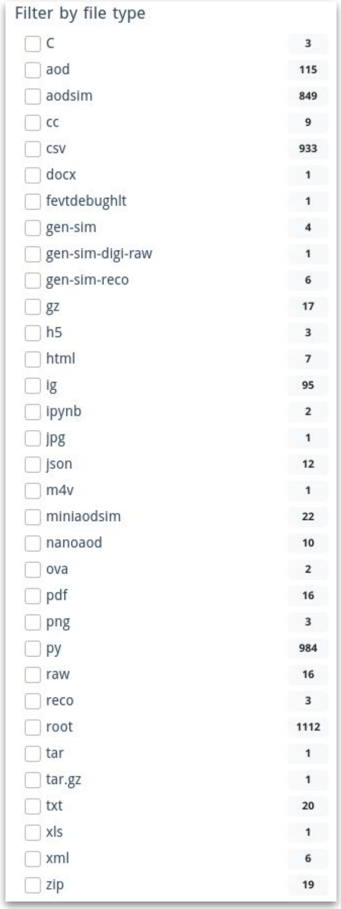
☐ Dataset	2185
☐ Collision	149
☐ Derived	1111
☐ Simulated	925
☐ Documentation	66
☐ About	9
☐ Activities	19
☐ Authors	5
☐ Guide	24
☐ Help	2
☐ Policy	6
☐ Report	1
☐ Environment	30
☐ Condition	9
☐ VM	16
☐ Validation	5
☐ Glossary	36
☐ News	14
☐ Software	44
☐ Analysis	18
☐ Framework	4
☐ Tool	16
☐ Validation	6
☐ Workflow	6
☐ Supplementaries	2703
☐ Configuration	58
☐ Configuration HLT	213
☐ Configuration LHE	242
☐ Configuration RECO	149
☐ Configuration SIM	313
☐ Luminosity	5
☐ Trigger	1723

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

Hierarchical data organisation

https://www.archiver-project.eu/deployment-scenarios-technical-summaries/cern-open-data

CERN Open Data: towards reuse and reproducibility



Filter by file type

□ C	3
□ aod	115
□ aodsim	849
□ cc	9
□ csv	933
□ docx	1
□ fevtdebughlt	1
□ gen-sim	4
□ gen-sim-digi-raw	1
□ gen-sim-reco	6
□ gz	17
□ h5	3
□ html	7
□ ig	95
□ ipynb	2
□ jpg	1
□ json	12
□ m4v	1
□ miniaodsim	22
□ nanoaod	10
□ ova	2
□ pdf	16
□ png	3
□ py	984
□ raw	16
□ reco	3
□ root	1112
□ tar	1
□ tar.gz	1
□ txt	20
□ xls	1
□ xml	6
□ zip	19

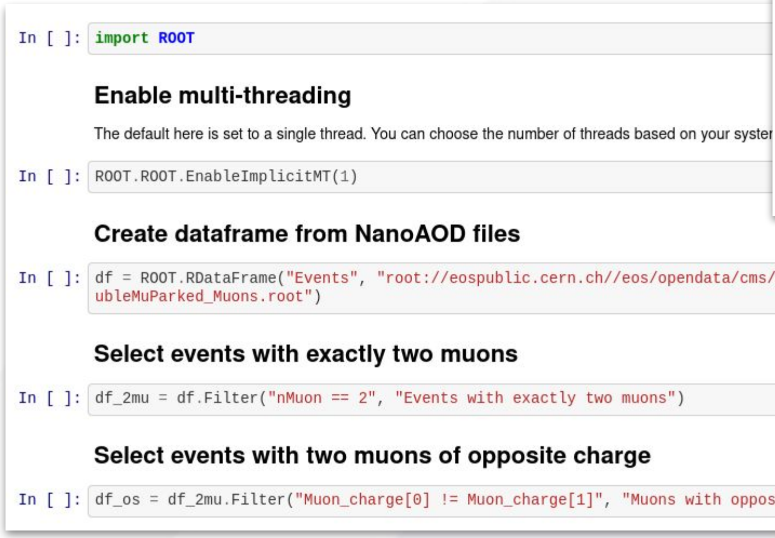
```
$ file mycode1.cc
mycode1.cc: C source, ASCII text

$ file mycode2.cc
mycode2.cc: Python script, ASCII text executable

$ file mydata.csv
mydata.csv: CSV text

$ csvlint mydata.csv
Record #15 has error: wrong number of fields in line
```

File content type verification



```
In [ ]: import ROOT
```

Enable multi-threading

The default here is set to a single thread. You can choose the number of threads based on your syster

```
In [ ]: ROOT.ROOT.EnableImplicitMT(1)
```

Create dataframe from NanoAOD files

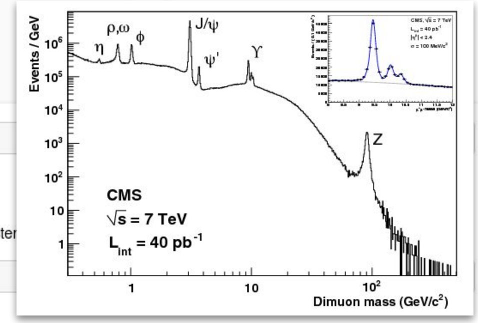
```
In [ ]: df = ROOT.RDataFrame("Events", "root://eospublic.cern.ch//eos/opendata/cms/derived-data/AOD2NanoAODOutreachTool/Run2012BC_Do
        ubleMuParked_Muons.root")
```

Select events with exactly two muons

```
In [ ]: df_2mu = df.Filter("nMuon == 2", "Events with exactly two muons")
```

Select events with two muons of opposite charge

```
In [ ]: df_os = df_2mu.Filter("Muon_charge[0] != Muon_charge[1]", "Muons with opposite charg
```

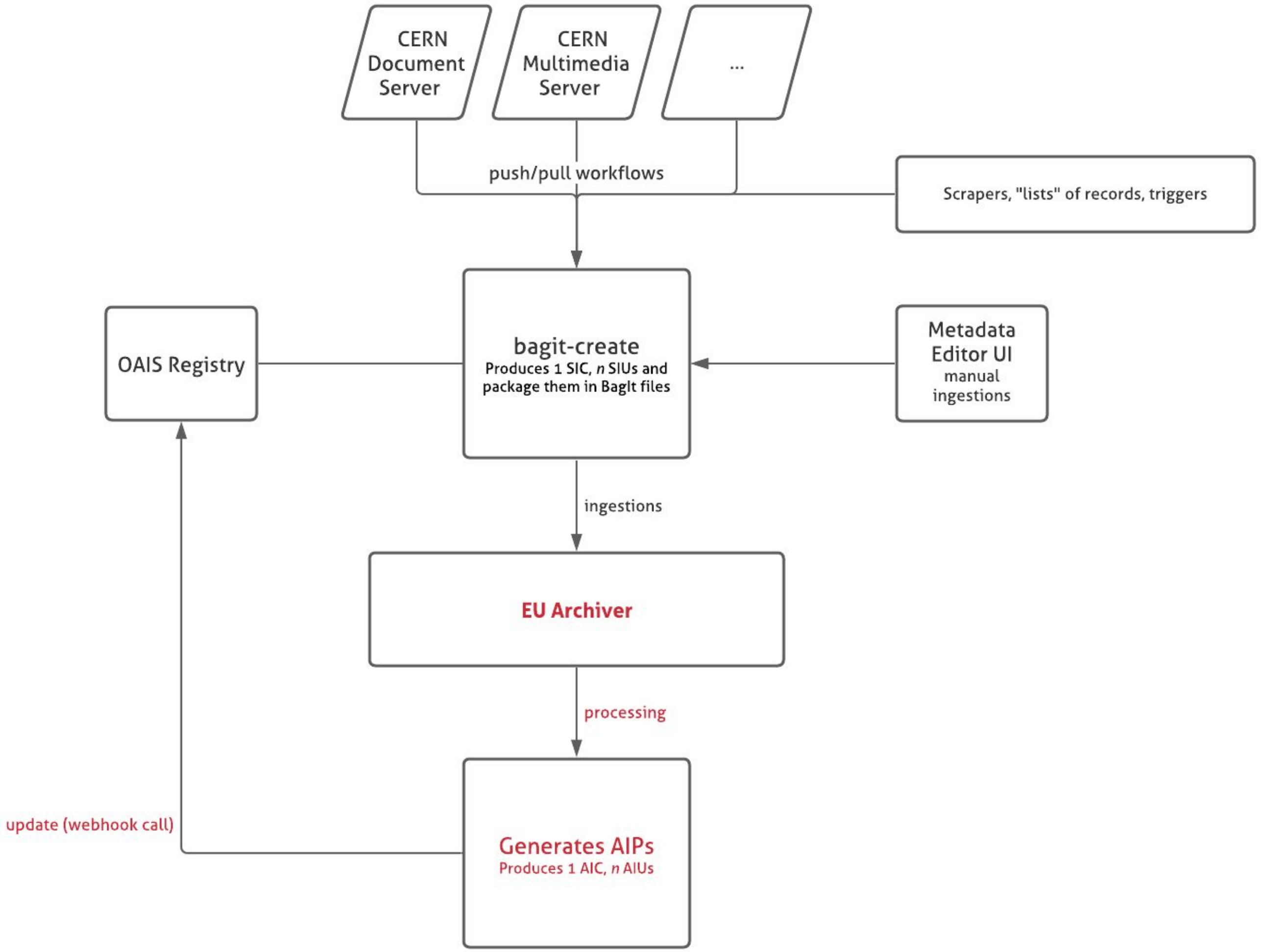


XRootD

Community-oriented data exposure

https://www.archiver-project.eu/deployment-scenarios-technical-summaries/cern-open-data

CERN Digital Memory: Archive for institutional data



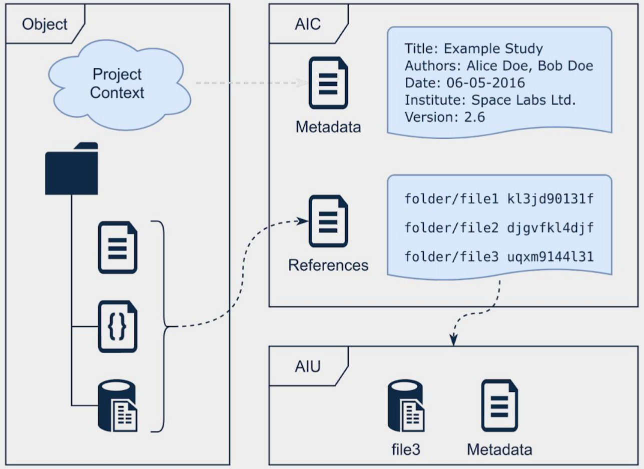
Diversity of input types: text, image, videos in many formats

Multiple pipelines for different sources

Challenges to address:
- duplication,
- authorships,
- integrity,
- versioning

CERN Digital Memory: AICs & AIUs

AIC versioning strategy



Title: Example Study
Authors: Alice Doe, Bob Doe
Date: 06-05-2016
Institute: Space Labs Ltd.
Version: 2.6

```
folder/file1 kl3jd90131f
folder/file2 djgvfkl4djf
folder/file3 uqxm9144l31
```

CERN OAIS Registry update with AIC IDs, possibly using webhook calls

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

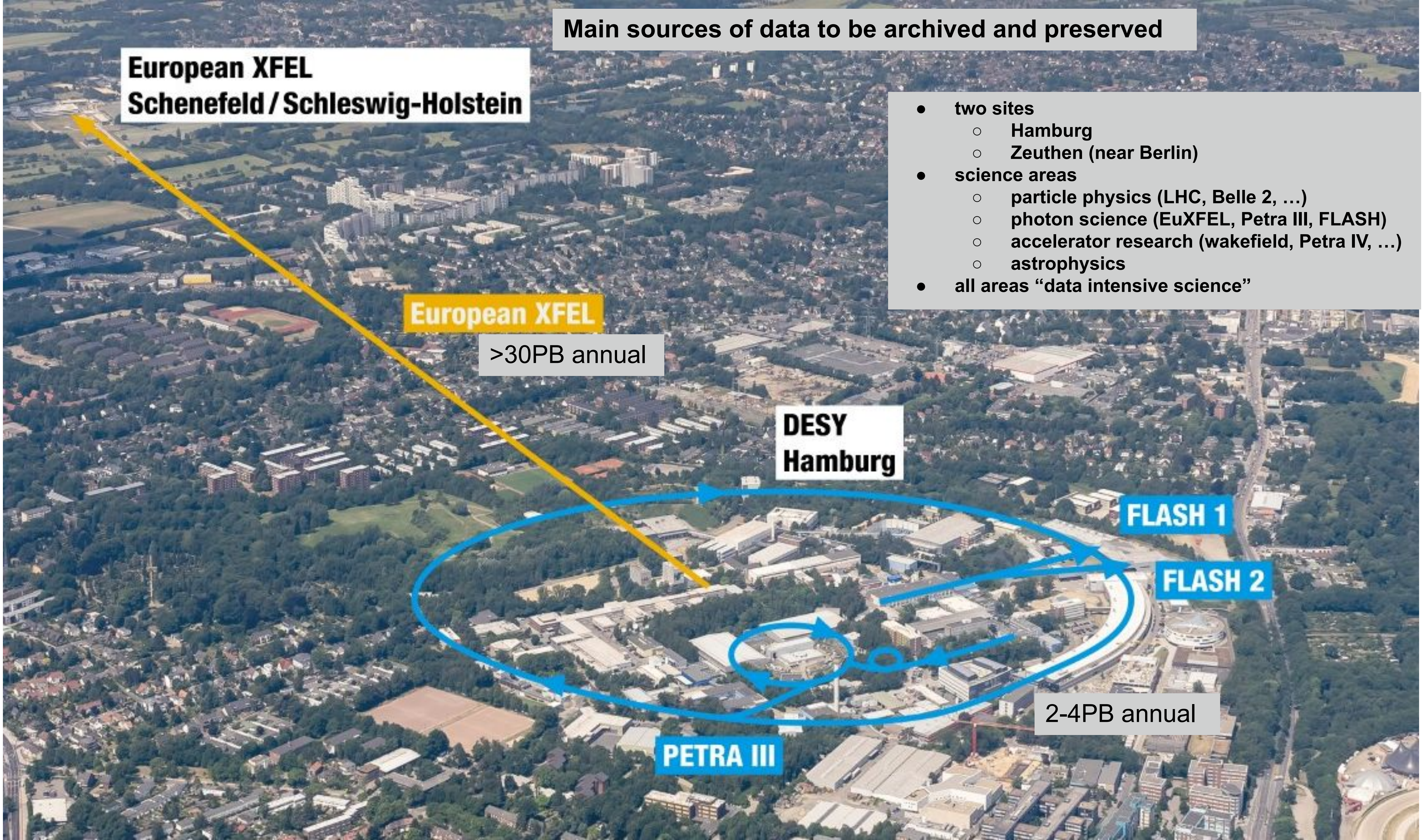
Ability to reconstruct original objects at any given point of time

DESY Requirements and Expectations

Sergey Yakubov, Martin Gasthuber

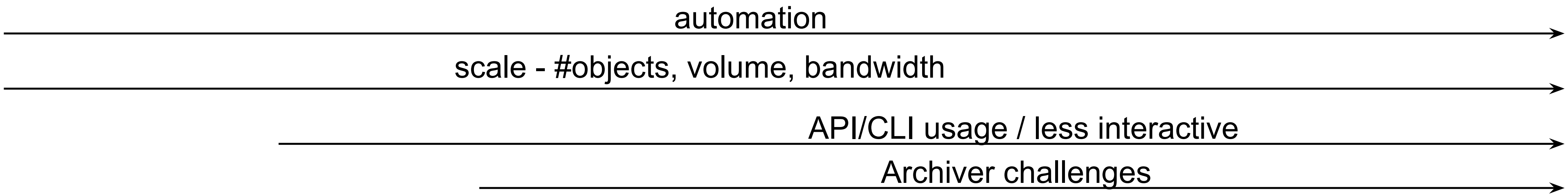
Main sources of data to be archived and preserved

- 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”



>30PB annual

2-4PB annual



automation

scale - #objects, volume, bandwidth

API/CLI usage / less interactive

Archiver challenges

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 !

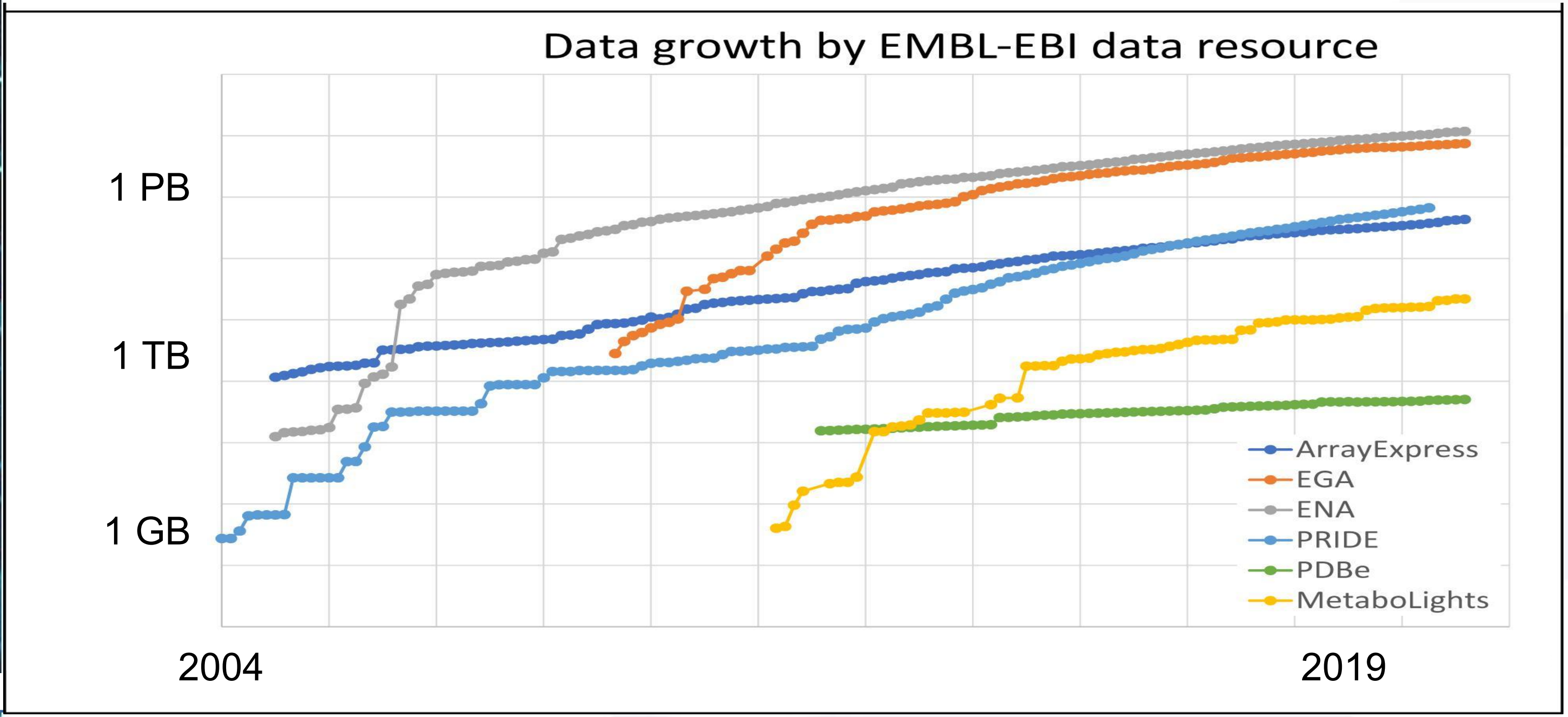
EMBL-EBI

EMBL-EBI Requirements and Expectations

Tony Wildish

EMBL-EBI

Increasing Data, Increasing Analysis



Data volume doubles every two years

- => half of our data will always be < 2 years old

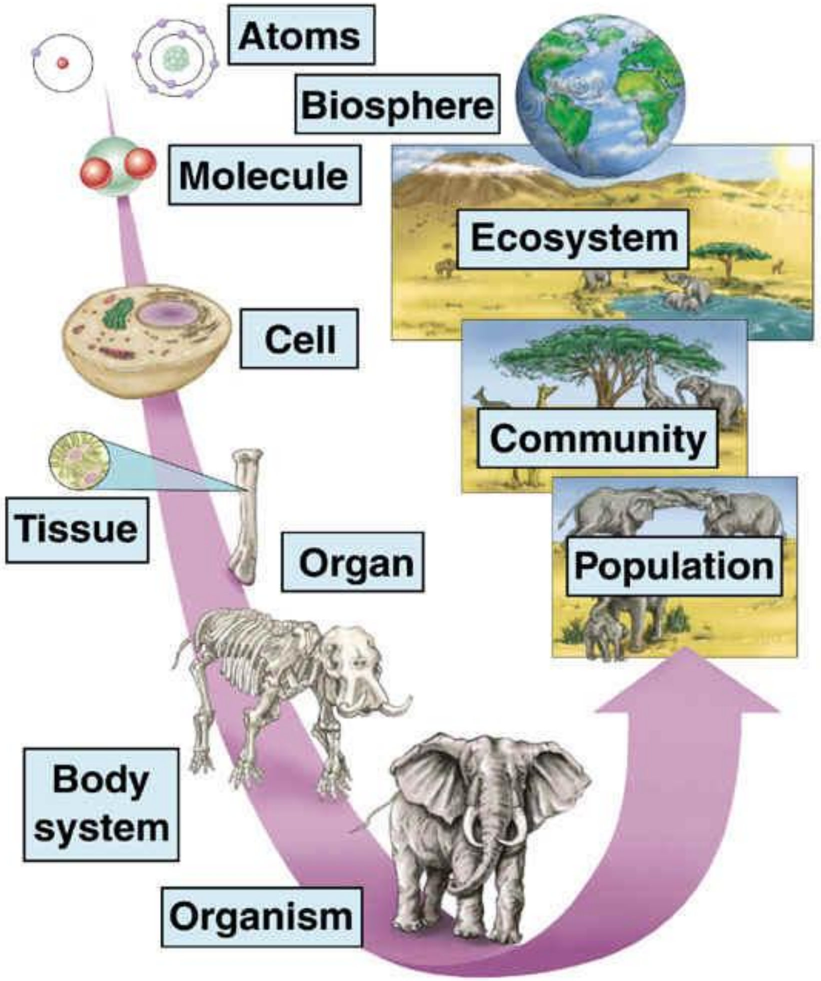
EBI data: *almost* Archival…

- 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!

Raven/Berg, Environment, 3/e
Figure 4.1



Harcourt, Inc.

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

PIC
port d'informació
científica

Use cases will be based on MAGIC Telescope data
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

Daniel López / IAC

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

ALBA

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”)

ARCHIVER

ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

Early Adopters Programme

Anna Manou - CERN

ARCHIVER - Archiving and Preservation for Research Environments project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824516.

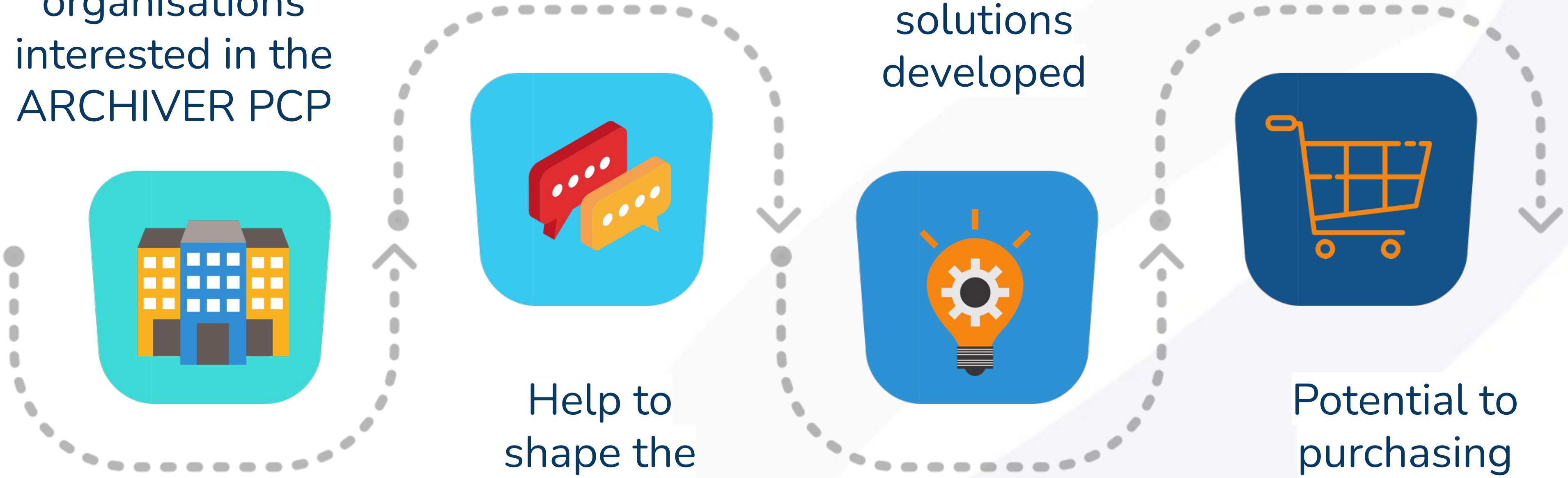
Early adopters Programme

WHAT?

WHY?

HOW?

WHAT?



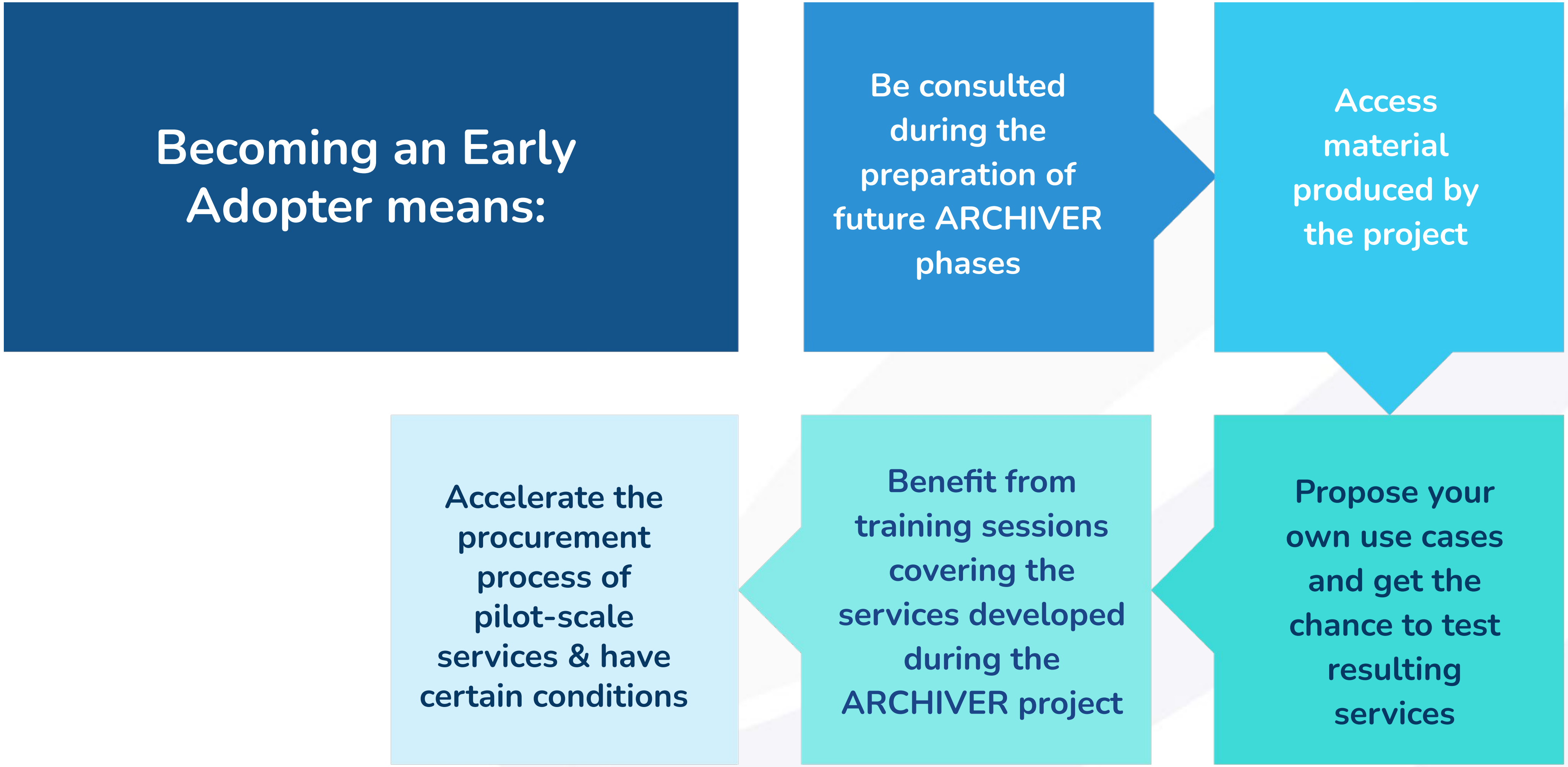
Public sector & not-for-profit organisations interested in the ARCHIVER PCP

Help to shape the R&D

Test the solutions developed

Potential to purchasing pilot-scale services

WHY?



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

CRG Centre for Genomic Regulation

IEO Istituto Europeo di Oncologia

SURF SARA

aarnet Australia's Academic and Research Network

KAUST

Universitet Stockholms

ALBA

SWITCH Serving Swiss Universities

Jisc

FMI Friedrich Miescher Institute for Biomedical Research

UKRI Science and Technology Facilities Council

ARDC Australian Research Data Commons

Use cases

IEO Istituto Europeo di Oncologia

Archival and accessibility of omics data

STOCKHOLMS UNIVERSITET

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

CRG Centre for Genomic Regulation

Archiving Genomic and Imaging Data

aarnet Australia's Academic and Research Network

Preserving Australia's digital research, education and cultural heritage

ARDC Australian Research Data Commons

Defining National Scale Data Archive Services

SCAN ME

https://archiver-project.eu/early-adopters-use-cases

HOW?

Are you part of a public sector research organisation with needs for standards-based, cost-effective data archiving and preservation services?

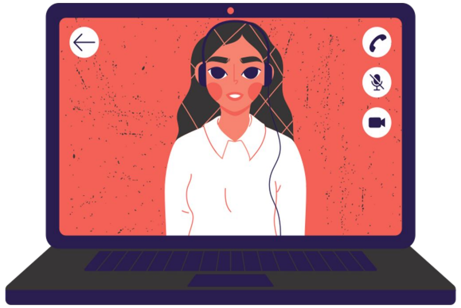
Are high ingest rates, data volumes at scale and long-term support important to you?

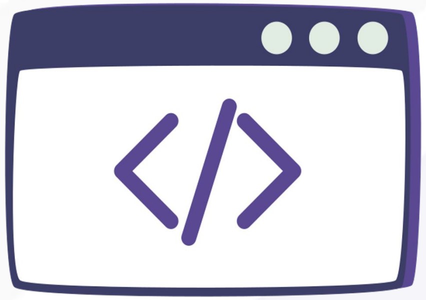
September 2021

Express your interest

SCAN ME

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





SCAN ME

SCAN ME

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

BREAK

Video: Outcomes of the design phase

ARCHIVER

ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

PROTOTYPE PHASE

AWARD CEREMONY

ARCHIVER - Archiving and Preservation for Research Environments project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824516.

ARCHIVER

ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

T-Systems International – GWDG – Onedata

T - Systems

ONEDATA

Open Telekom Cloud

GWDG
Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen

ARCHIVER - Archiving and Preservation for Research Environments project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824516.

ARCHIVER
Cloud-agnostic approach for research
GWDG
Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen
ONEDATA
T··Systems·
0:01 / 2:29
CC

https://tinyurl.com/y23gskwj

ARCHIVER

ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

ARCHIVING AND PRESERVING TO DISCOVER

PROTOTYPE PHASE

Jurry de la Mar

T · · Systems ·

ARCHIVING AND PRESERVATION FOR RESEARCH

MISSION 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

Prof. Dr. Philipp Wieder

Research Data and Preservation Expert, GWDG

Jurry de la Mar

Science and Research Expert, T-Systems

Dietrich Singh

Cloud Expert, T-Systems

Lukasz Dutka

Research Data Expert, Onedata

Prof. Dr. Ramin Yahyapour

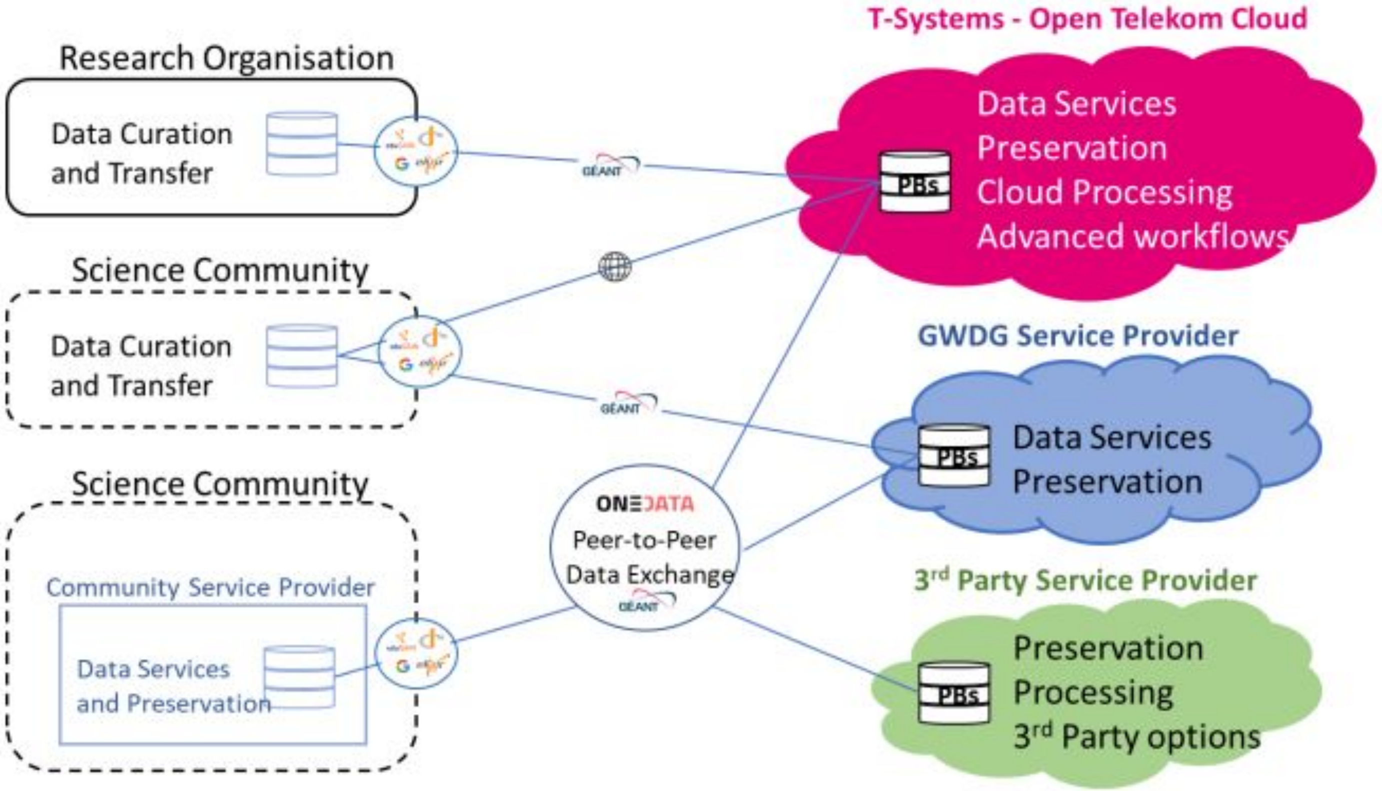
Research Data and IT Expert, GWDG

Bartosz Kryza

Distributed Data Expert, Onedata

DESIGN: OPEN-SOURCE AND CLOUD-AGNOSTIC

- 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



Research Organisation

Data Curation and Transfer

Science Community

Data Curation and Transfer

Science Community

Community Service Provider

Data Services and Preservation

ONEDATA Peer-to-Peer Data Exchange

GÉANT

T-Systems - Open Telekom Cloud

PBs

Data Services
Preservation
Cloud Processing
Advanced workflows

GWDG Service Provider

PBs

Data Services
Preservation

3rd Party Service Provider

PBs

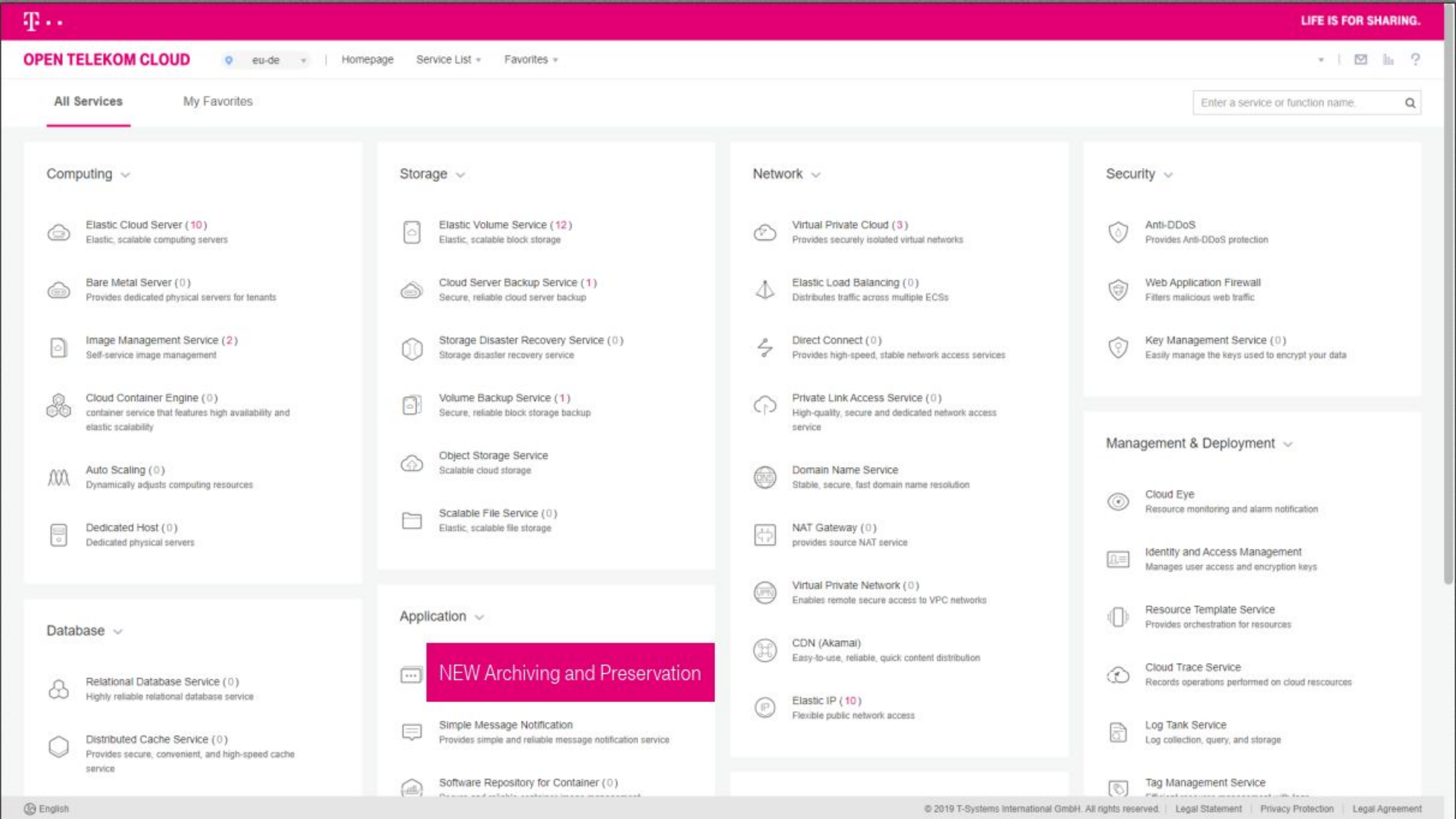
Preservation
Processing
3rd Party options

BPMN = Business Process Model and Notation Standard

PROTOTYPE FOCUS

Functionality	Prototype Phase	Pilot Phase
Sustainability	X	X
Hybrid Deployments	X	X
Data redundancy	X	X
Data-at-rest encryption	X	X
Data immutability	X	X
Extended KPI monitoring	X	X
CTS/ISO certification	Preparation	X
Intelligent caching		X
Cloud UI integration		X

The proof of the pudding is in the eating!



LIFE IS FOR SHARING.

OPEN TELEKOM CLOUD

All Services My Favorites

Computing

- Elastic Cloud Server (10)
  Elastic, scalable computing servers
- Bare Metal Server (0)
  Provides dedicated physical servers for tenants
- Image Management Service (2)
  Self-service image management
- Cloud Container Engine (0)
  container service that features high availability and elastic scalability
- Auto Scaling (0)
  Dynamically adjusts computing resources
- Dedicated Host (0)
  Dedicated physical servers

Database

- Relational Database Service (0)
  Highly reliable relational database service
- Distributed Cache Service (0)
  Provides secure, convenient, and high-speed cache service

Storage

- Elastic Volume Service (12)
  Elastic, scalable block storage
- Cloud Server Backup Service (1)
  Secure, reliable cloud server backup
- Storage Disaster Recovery Service (0)
  Storage disaster recovery service
- Volume Backup Service (1)
  Secure, reliable block storage backup
- Object Storage Service
  Scalable cloud storage
- Scalable File Service (0)
  Elastic, scalable file storage

Application

NEW Archiving and Preservation

- Simple Message Notification
  Provides simple and reliable message notification service
- Software Repository for Container (0)

Network

- Virtual Private Cloud (3)
  Provides securely isolated virtual networks
- Elastic Load Balancing (0)
  Distributes traffic across multiple ECSs
- Direct Connect (0)
  Provides high-speed, stable network access services
- Private Link Access Service (0)
  High-quality, secure and dedicated network access service
- Domain Name Service
  Stable, secure, fast domain name resolution
- NAT Gateway (0)
  provides source NAT service
- Virtual Private Network (0)
  Enables remote secure access to VPC networks
- CDN (Akamai)
  Easy-to-use, reliable, quick content distribution
- Elastic IP (10)
  Flexible public network access

Security

- Anti-DDoS
  Provides Anti-DDoS protection
- Web Application Firewall
  Filters malicious web traffic
- Key Management Service (0)
  Easily manage the keys used to encrypt your data

Management & Deployment

- Cloud Eye
  Resource monitoring and alarm notification
- Identity and Access Management
  Manages user access and encryption keys
- Resource Template Service
  Provides orchestration for resources
- Cloud Trace Service
  Records operations performed on cloud rescources
- Log Tank Service
  Log collection, query, and storage
- Tag Management Service

English

© 2019 T-Systems International GmbH. All rights reserved. | Legal Statement | Privacy Protection | Legal Agreement

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

FREEMIUM

PREMIUM

COLLABORATE

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.

ARCHIVER

ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

Libnova – CSIC – University of Barcelona – Giaretta Associates - AWS - Voxility

ARCHIVER - Archiving and Preservation for Research Environments project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824516.

https://tinyurl.com/yxss3py3

Building the next generation **Research Data Management** solution

libnova

the most advanced digital preservation platform

Consortium

libnova
the most advanced digital preservation platform

LIBERTAS PERFVNDET · OMNIA LVCE
UNIVERSITAT DE BARCELONA

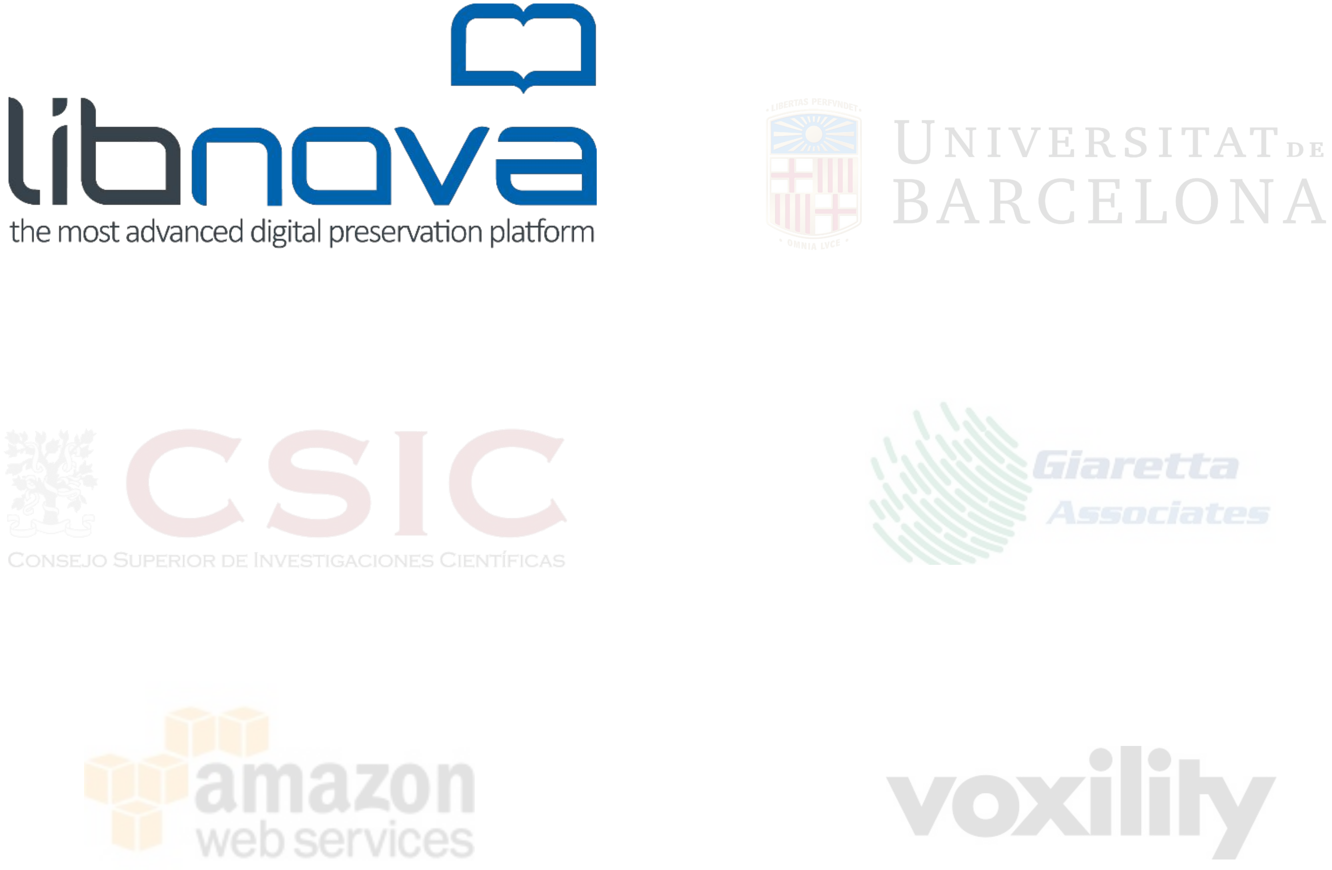
CSIC
CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

Giaretta Associates

amazon web services

voxility

Consortium


libnova
the most advanced digital preservation platform
UNIVERSITAT DE BARCELONA
CSIC
CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS
Giaretta Associates
amazon web services
voxility

- **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.

Consortium

libnova
the most advanced digital preservation platform

UNIVERSITAT DE BARCELONA

CSIC
CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

Giaretta Associates

amazon web services

voxility

- 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.

Consortium

libnova
the most advanced digital preservation platform

UNIVERSITAT DE BARCELONA

CSIC
CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

Giaretta Associates

amazon web services

voxility

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.

Consortium

libnova
the most advanced digital preservation platform
Universitat de Barcelona
CSIC
Consejo Superior de Investigaciones Científicas
Giaretta Associates
amazon web services
voxility

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.

Consortium

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.

Consortium

libnova
the most advanced digital preservation platform

UNIVERSITAT DE BARCELONA

CSIC
CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

Giaretta Associates

amazon web services

voxility

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.

Consortium

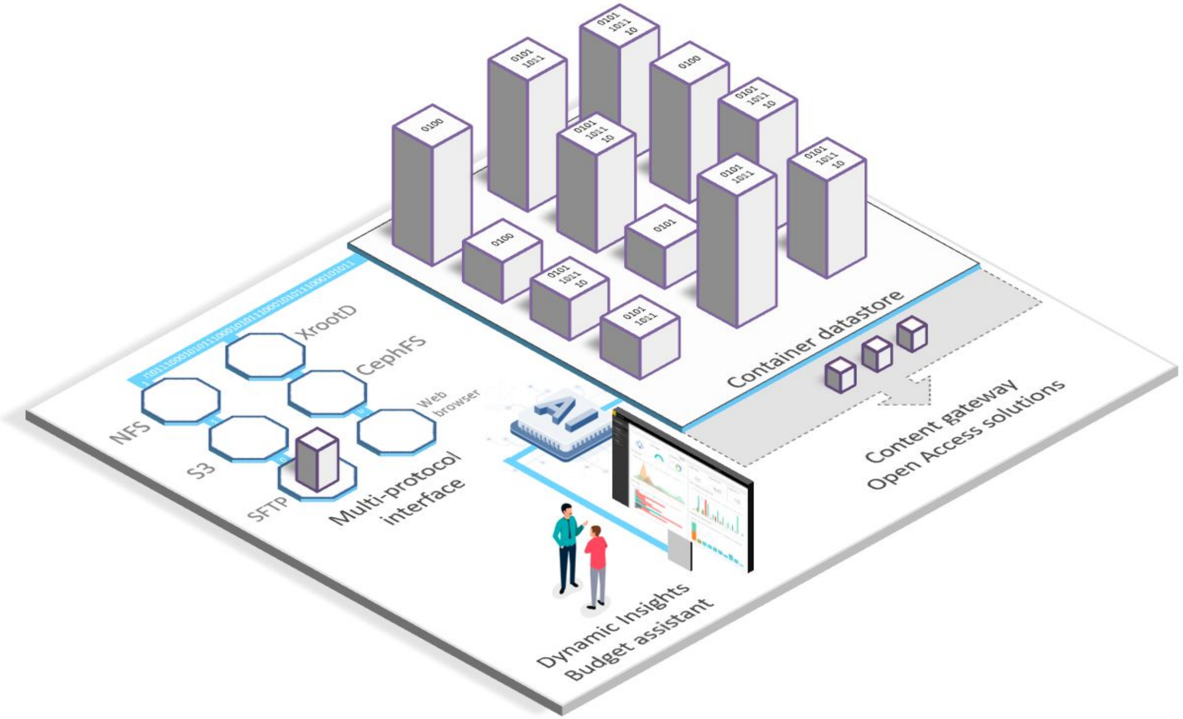
- Multi-petabyte scale with the **CSIC**’s vast experience on supercomputing and large-scale infrastructures plus **Amazon Web Services** and **Voxility Infrastructure**.

- 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.

Solution



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.

Solution

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 standandars: 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.

the most advanced digital preservation platform

Contact LIBNOVA:
contact@libnova.com

Contact me:
a.guillermo@libnova.com

the most advanced digital preservation platform

ARCHIVER

HELIX NEBULA

ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

Arkivum - Google

arkivum

Bringing archived data to life

Google Cloud

ARCHIVER - Archiving and Preservation for Research Environments project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824516.

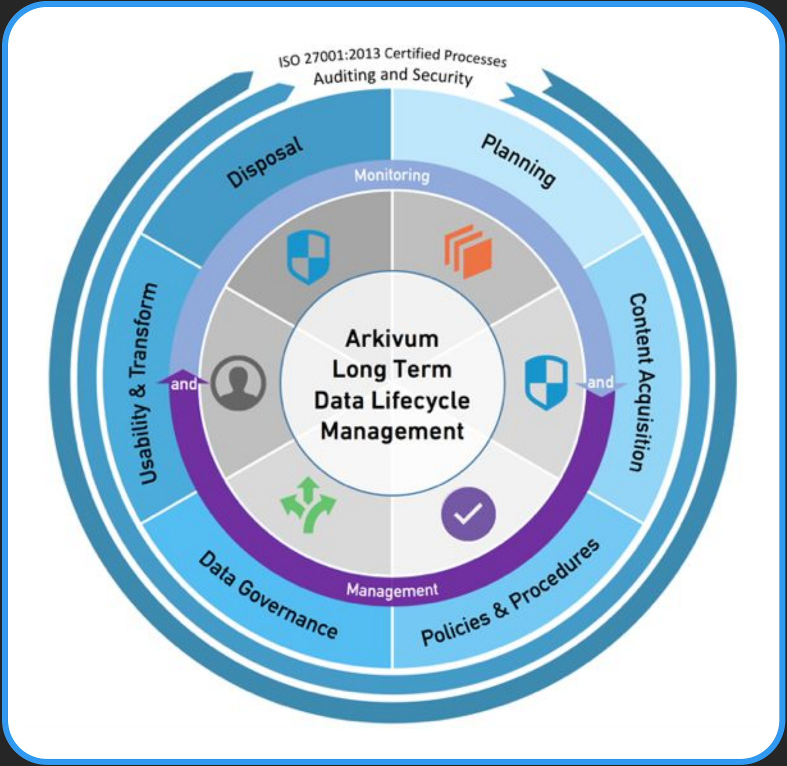
https://tinyurl.com/y4ycaqpa

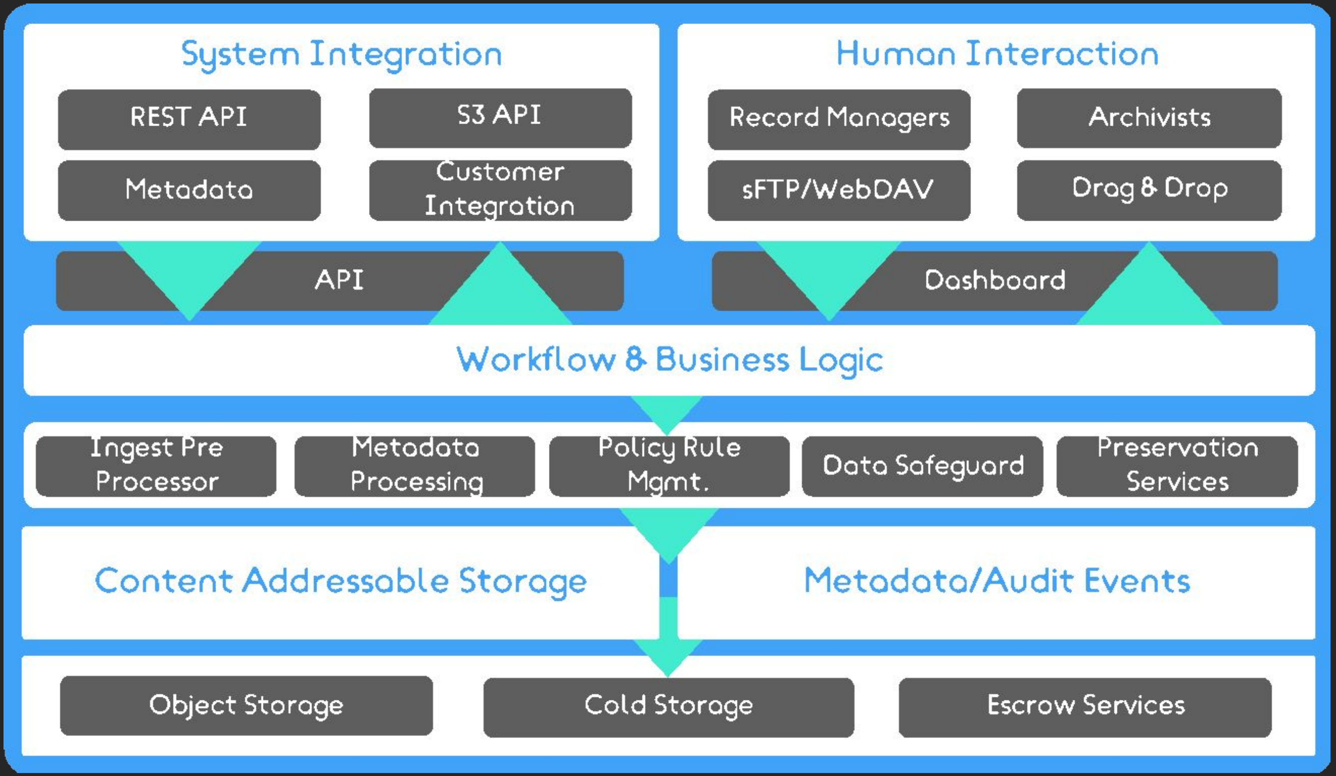
arkivum

Bringing archived data to life

ARCHIVER
Arkivum and Google solution
Phase 2: Prototype

Arkivum Perpetua: Cloud Hosted Digital Preservation and Archiving





System Integration

REST API

S3 API

Metadata

Customer Integration

Human Interaction

Record Managers

Archivists

sFTP/WebDAV

Drag & Drop

API

Dashboard

Workflow & Business Logic

Ingest Pre Processor

Metadata Processing

Policy Rule Mgmt.

Data Safeguard

Preservation Services

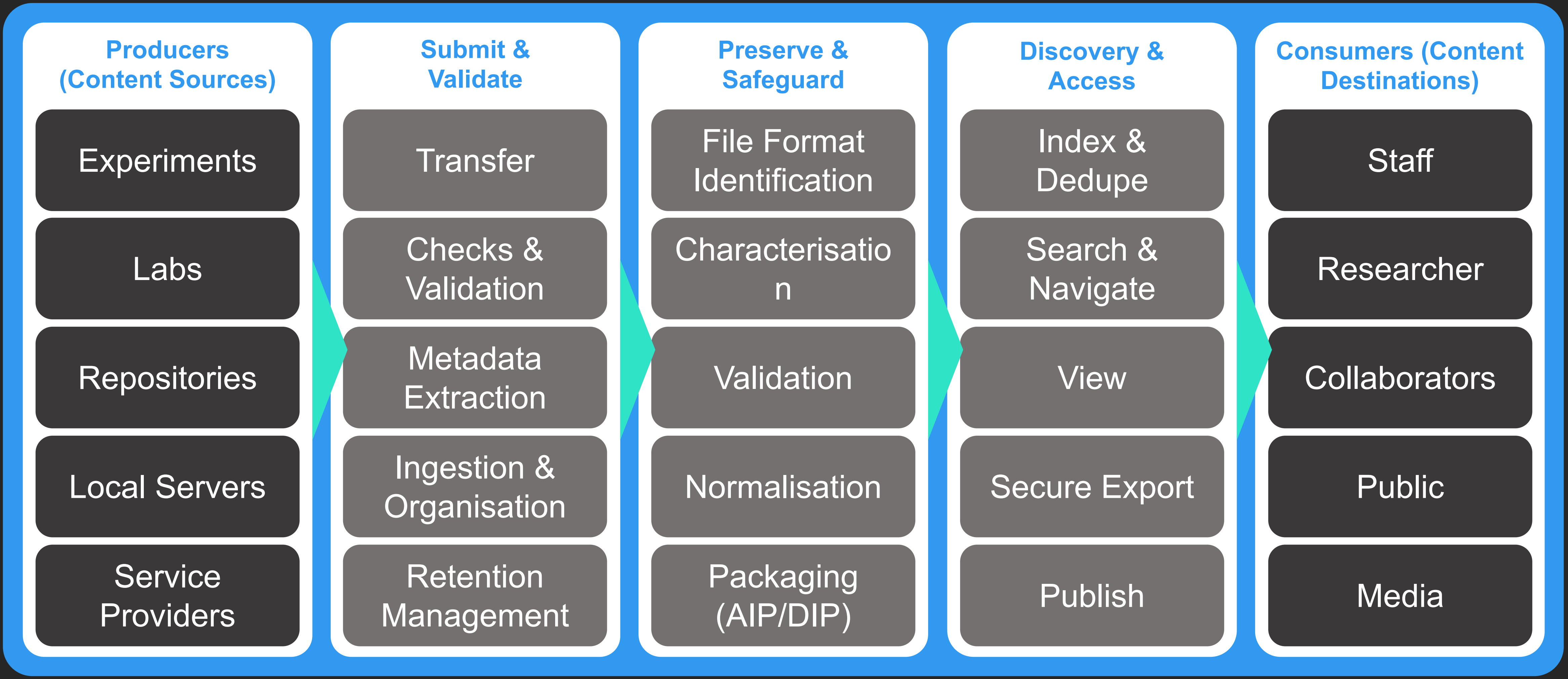
Content Addressable Storage

Metadata/Audit Events

Object Storage

Cold Storage

Escrow Services

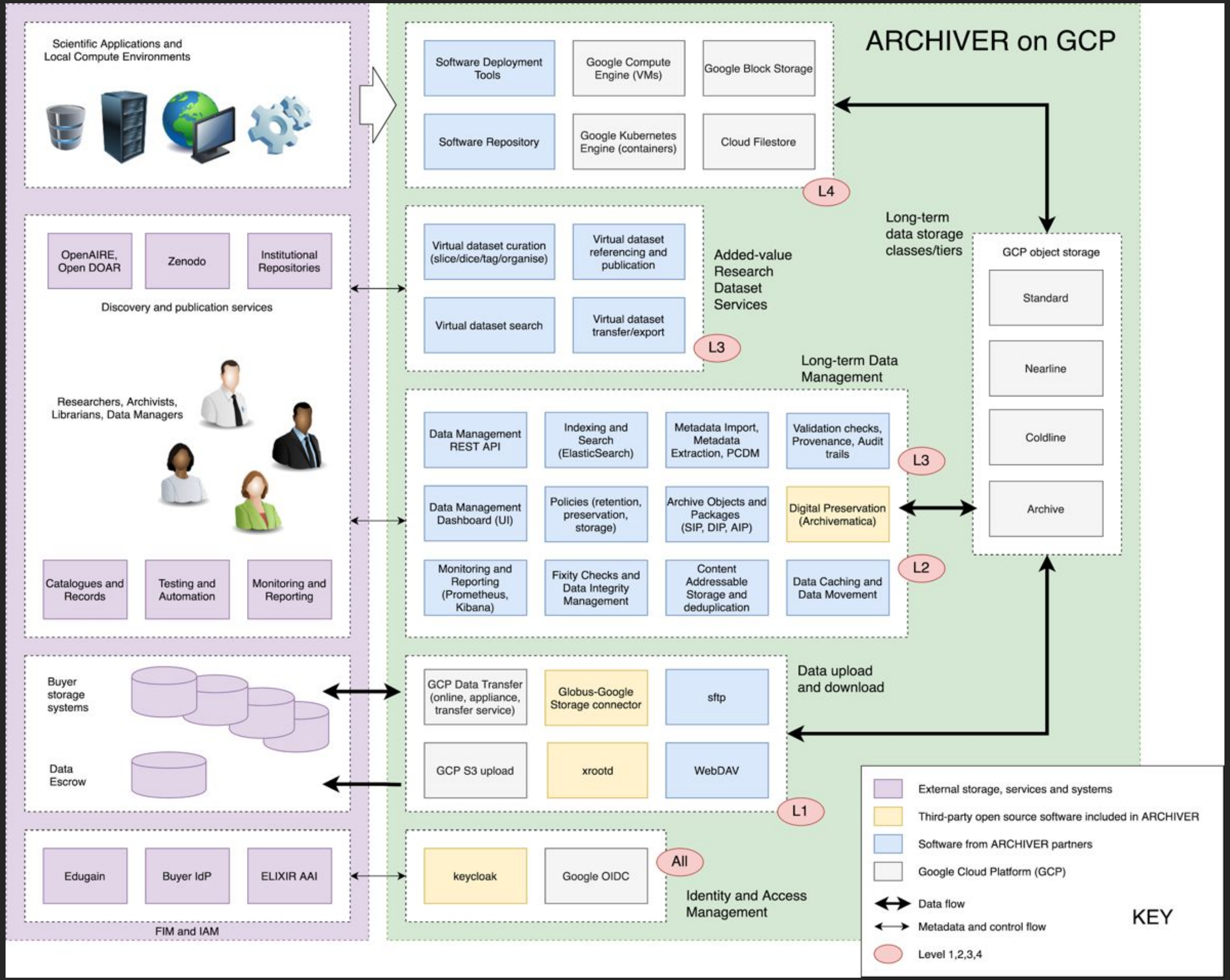


Producers (Content Sources)	Submit & Validate	Preserve & Safeguard	Discovery & Access	Consumers (Content Destinations)
Experiments	Transfer	File Format Identification	Index & Dedupe	Staff
Labs	Checks & Validation	Characterisation	Search & Navigate	Researcher
Repositories	Metadata Extraction	Validation	View	Collaborators
Local Servers	Ingestion & Organisation	Normalisation	Secure Export	Public
Service Providers	Retention Management	Packaging (AIP/DIP)	Publish	Media

OAIS, TDR, Core Trust Seal, DPC RAM, Nestor

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



ARCHIVER on GCP

Scientific Applications and Local Compute Environments

Software Deployment Tools
Google Compute Engine (VMs)
Google Block Storage
Software Repository
Google Kubernetes Engine (containers)
Cloud Filestore
L4

OpenAIRE, Open DOAR
Zenodo
Institutional Repositories
Discovery and publication services

Virtual dataset curation (slice/dice/tag/organise)
Virtual dataset referencing and publication
Virtual dataset search
Virtual dataset transfer/export
L3
Added-value Research Dataset Services

Long-term data storage classes/tiers
GCP object storage
Standard
Nearline
Coldline
Archive

Researchers, Archivists, Librarians, Data Managers

Long-term Data Management
Data Management REST API
Indexing and Search (ElasticSearch)
Metadata Import, Metadata Extraction, PCDM
Validation checks, Provenance, Audit trails
L3
Data Management Dashboard (UI)
Policies (retention, preservation, storage)
Archive Objects and Packages (SIP, DIP, AIP)
Digital Preservation (Archivematica)
Monitoring and Reporting (Prometheus, Kibana)
Fixity Checks and Data Integrity Management
Content Addressable Storage and deduplication
Data Caching and Data Movement
L2

Catalogues and Records
Testing and Automation
Monitoring and Reporting

Buyer storage systems
Data Escrow

GCP Data Transfer (online, appliance, transfer service)
Globus-Google Storage connector
sftp
GCP S3 upload
xrootd
WebDAV
L1
Data upload and download

Edugain
Buyer IdP
ELIXIR AAI
FIM and IAM

keycloak
Google OIDC
All
Identity and Access Management

KEY

▨	External storage, services and systems
▨	Third-party open source software included in ARCHIVER
▨	Software from ARCHIVER partners
□	Google Cloud Platform (GCP)
⟷	Data flow
↔	Metadata and control flow
○	Level 1,2,3,4

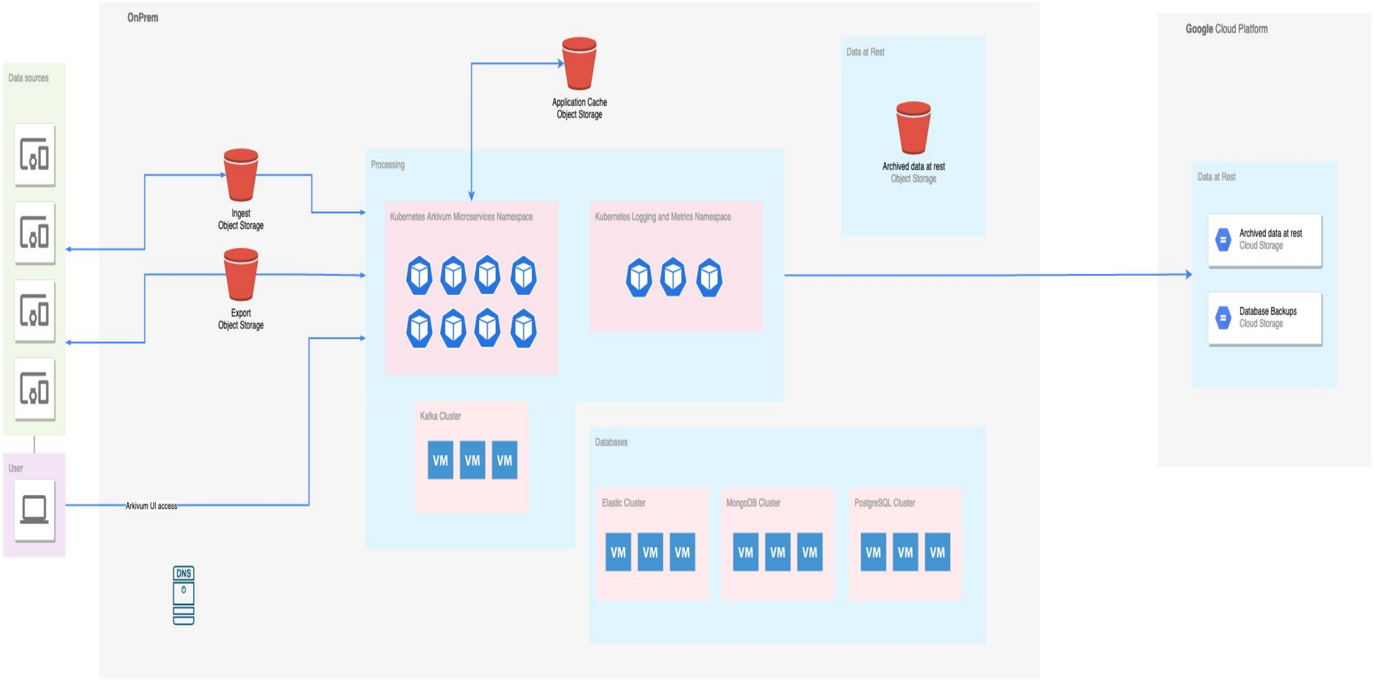
Google Cloud Platform: PB Scale Storage, Compute and Networking

- Google Object Storage
- Google File Storage
- Google Compute Engine
- Google Kubernetes Engine

- Google Operations
- Google Security
- High speed network
- GÉANT
- 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



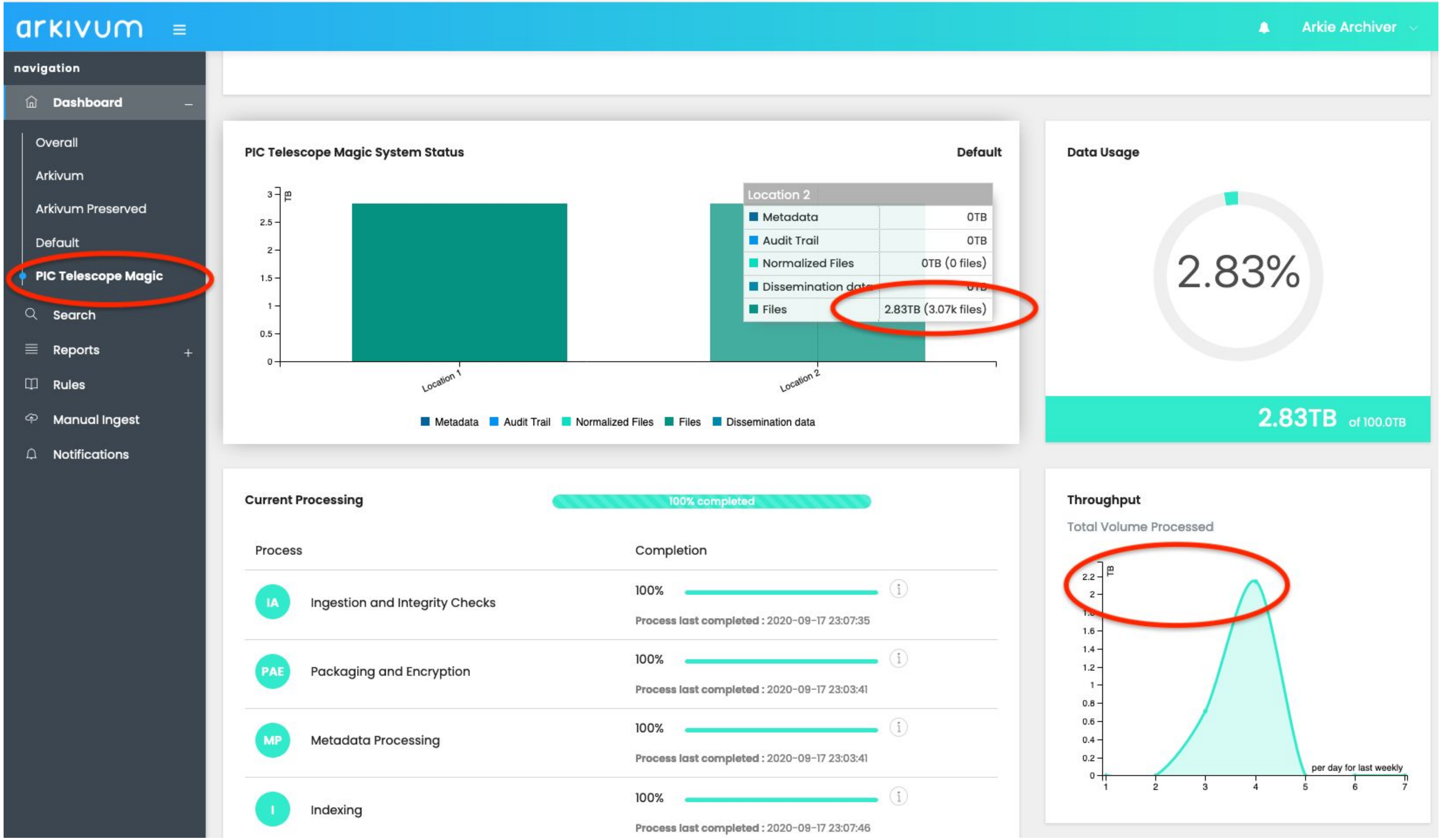
METS Metadata Encoding & Transmission Standard Official Web Site

PREMIS PRESERVATION METADATA MAINTENANCE ACTIVITY

DURASPACE™
Portland Common Data Model

kubernetes docker Anthos mongoDB APACHE kafka® KEYCLOAK elasticsearch

Pilot: Long Term Digital Preservation Hosted On GCP



PIC Telescope Magic System Status — Default

Location 2	
Metadata	0TB
Audit Trail	0TB
Normalized Files	0TB (0 files)
Dissemination data	0TB
Files	2.83TB (3.07k files)

Data Usage

2.83%

2.83TB of 100.0TB

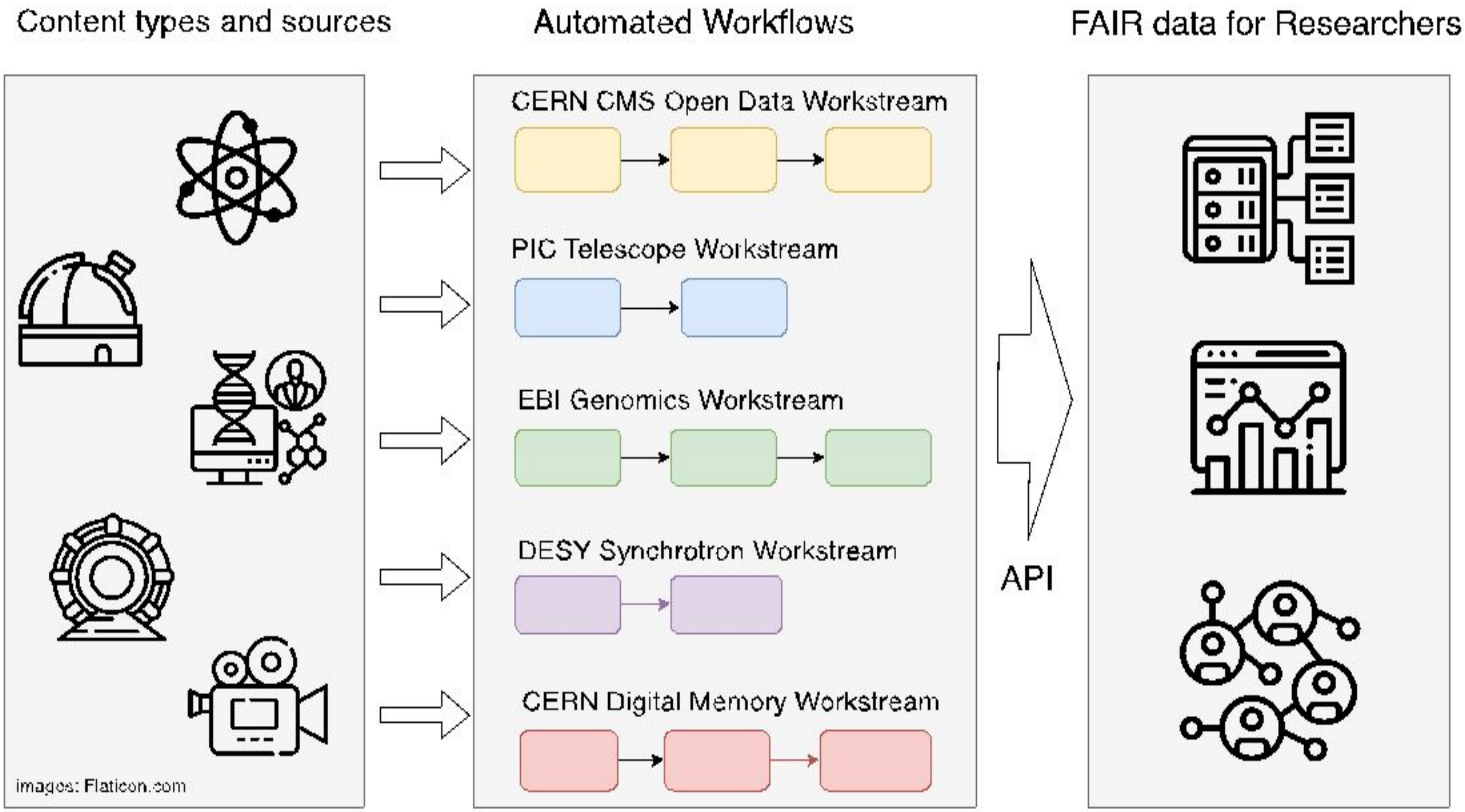
Current Processing — 100% completed

Process	Completion
IA Ingestion and Integrity Checks	100% Process last completed : 2020-09-17 23:07:35
PAE Packaging and Encryption	100% Process last completed : 2020-09-17 23:03:41
MP Metadata Processing	100% Process last completed : 2020-09-17 23:03:41
I Indexing	100% Process last completed : 2020-09-17 23:07:46

Throughput

Total Volume Processed

Prototype: Factories for LTDP in Large Scale Science



Content types and sources

Automated Workflows

FAIR data for Researchers

CERN CMS Open Data Workstream

PIC Telescope Workstream

EBI Genomics Workstream

DESY Synchrotron Workstream

CERN Digital Memory Workstream

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

arkivum

Bringing archived data to life

Thank you

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

www.arkivum.com

Find us on LinkedIn or on Twitter @Arkivum

London Office

Top Floor, The Walbrook Building
25 Walbrook, London EC4N 8AF UK
T: +44 (0)1249 40 50 60
E: hello@arkivum.com

Reading Office

Landmark, 450 Brook Drive, Green Park
Reading, Berkshire RG2 6UU UK
T: +44 (0)1249 40 50 60
E: hello@arkivum.com

ARCHIVER

ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

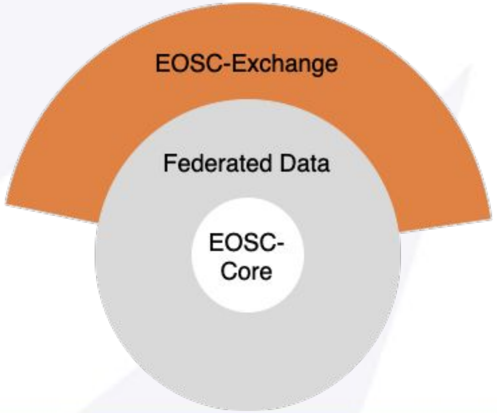
Feedback Session

Marion Devouassoux - CERN

ARCHIVER - Archiving and Preservation for Research Environments project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824516.

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?



Mentimeter

Go to menti.com

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