ARCHIVER

ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

PILOT END PHASE EVENT

13 June 2022
Moderator: Sara Pittonet (Trust-IT)
contact: s.pittonet@trust-itservices.com

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.

ARCHIVER finalist for the Digital Preservation Awards 2022

Digital Preservation Awards 2022 Finalists
Announced on 9th June, DPC Member voting open until 4th July

ARCHIVER work recognised for the International Council on Archives Award for Collaboration and Cooperation

which celebrates significant collaboration across institutional, professional, sectoral and geographical boundaries which have had a demonstrable and positive impact on digital preservation.

Event Outline

09.00 am - 09.10 am: Project Overview - João Fernandes (CERN)

09.10 am - 09.35 am: Pilot Phase - Buyers Group use cases (CERN, DESY, EMBL-EBI, PIC)

09.35 am - 09.45 am: Early Adopter Use Case: ECMWF Open Data (TBC)

09.45 - 09.55: Break

09.55 am- 10.25 am: Presentation from Libnova consortium (Antonio Martinez)

10.25 am - 10.55 am: Presentation from Arkivum consortium (Matthew Addis and Tom Lynam)

10.55 am - 11.00 am: Closing remarks - João Fernandes (CERN)

House Keeping

- This event is being recorded in its entirety. A link to the full recordings will be shared with participants afterwards.
- Post your questions in the chat. Q&A sessions are foreseen before the break and at the end of each consortium presentation.
- Please do not activate your microphone and videos unless the host gives you permission.

ARCHIVER

ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

Project Overview

Pilot End Phase

João Fernandes (CERN)
Contact: joao.fernandes@cern.ch

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 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)*

European Commission

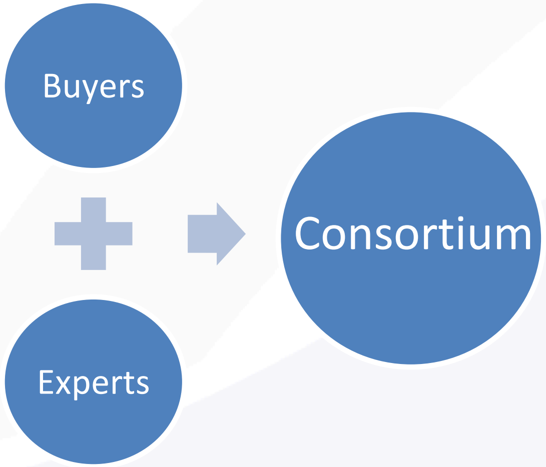
EUROPEAN OPEN SCIENCE CLOUD

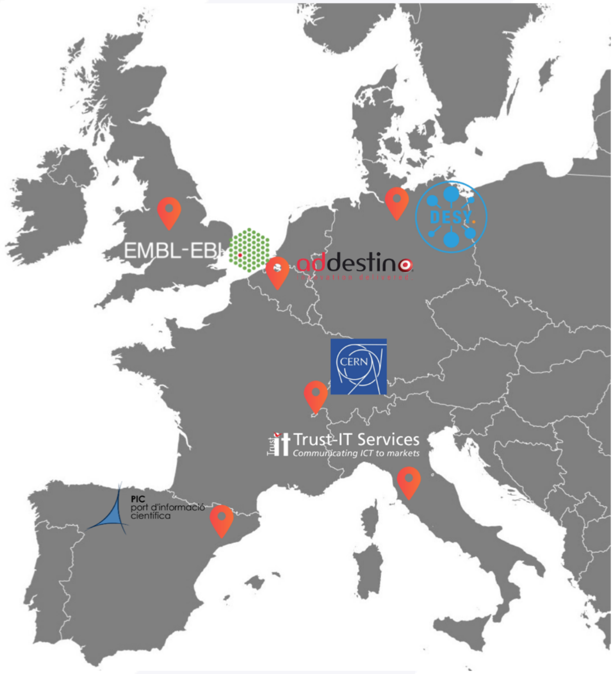
CERN DESY. EMBL-EBI PIC port d'informació científica

Buyers Group (BG) - Public organisations committing funds to contribute to a joint-R&D-procurement, research data use cases and R&D testing effort

addestino innovation delivered. Trust it

Experts - Partner organisations bringing expertise in requirement assessment and promotion activities



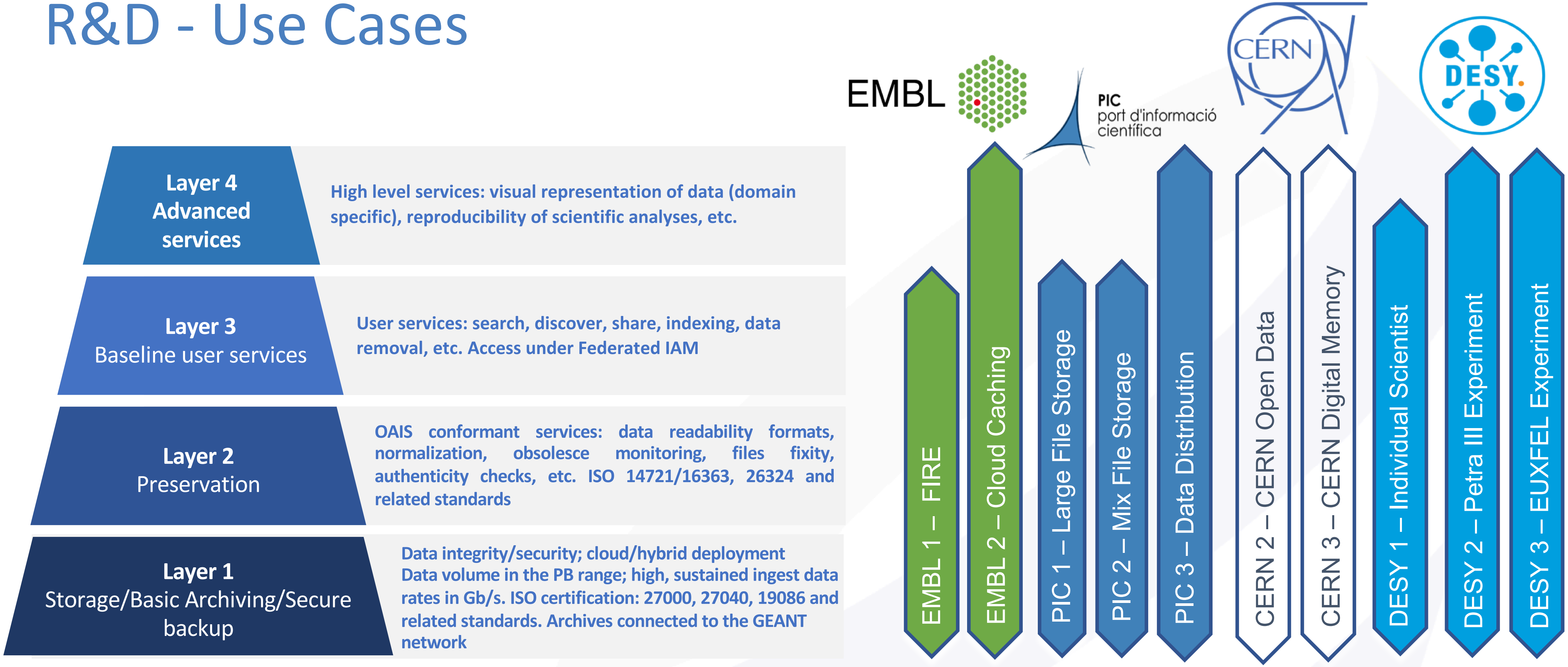


Progress Beyond the state of the art

Scientific Data Repositories before ARCHIVER		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 published	→	*Best practices: FAIR, TRUST, DPC RAM*
Technology lock-ins concerns (tape), *Business Continuity/Disaster Recovery (COVID-19)*	→	*Promote FOSS, open standards & demonstrate exit strategies*
Fragmentation across scientific disciplines & countries	→	*Pan-European: resulting services to be available in the EOSC portfolio*
Cost underestimation at the planning phase	→	*Cost-effective and environmentally sustainable model adapted to public research*

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

R&D - Use Cases



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 in Gb/s. ISO certification: 27000, 27040, 19086 and related standards. Archives connected to the GEANT network

EMBL
PIC
port d'informació científica
CERN
DESY.
EMBL 1 – FIRE
EMBL 2 – Cloud Caching
PIC 1 – Large File Storage
PIC 2 – Mix File Storage
PIC 3 – Data Distribution
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 "current state of the art" report in the context of the EOSC: https://doi.org/10.5281/zenodo.3618215

R&D Competitive Execution

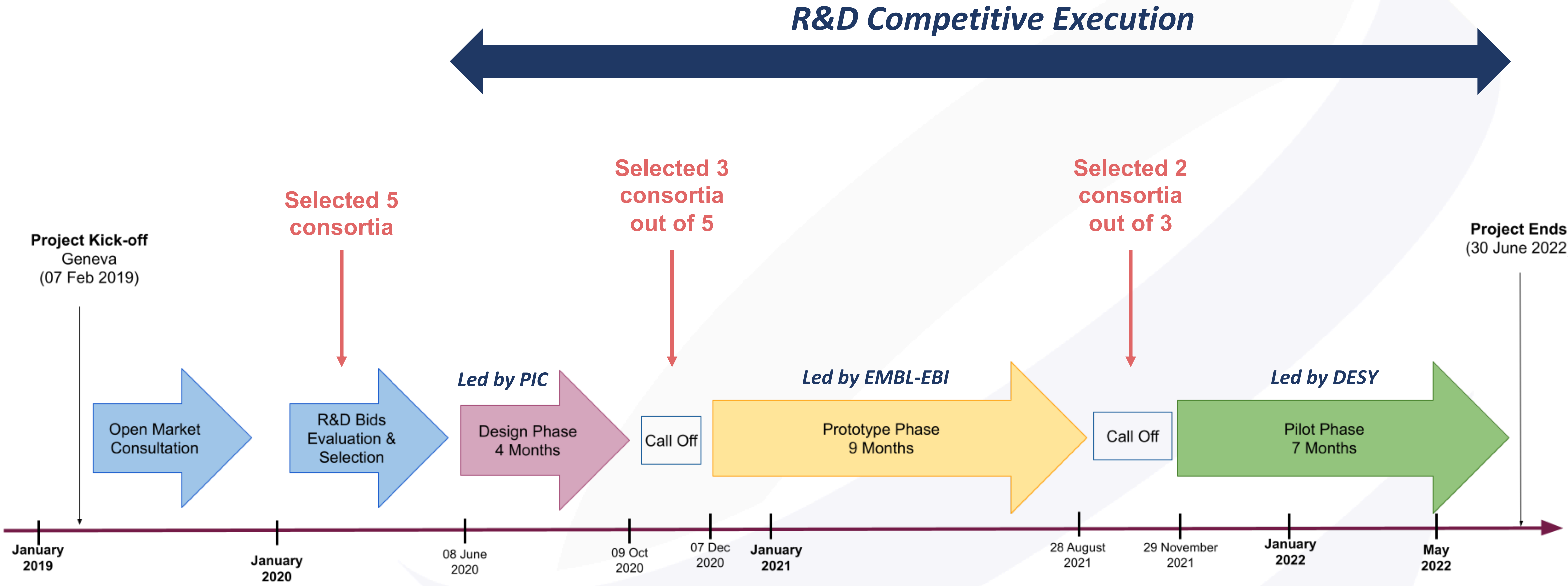
ARCHIVER followed an R&D execution on three phases with multiple competing consortia:

- **Phase 1 - Solution Design**
    - 5 selected consortia: https://archiver-project.eu/design-phase-award
    - Design architecture and technical components
    - Assessment & Selection process for consortia to proceed to the subsequent project phase

- **Phase 2 - Prototype Development**
    - 3 selected Designs: https://archiver-project.eu/prototype-phase-award
    - Prototypes building based on the design architectures
    - Functional testing and validation by the Buyers
    - Assessment and selection of Prototypes: Eligibility criteria to promote Prototypes to Pilots

- **Phase 3 - Pilot Deployment**
    - 2 selected Prototypes: https://archiver-project.eu/pilot-phase-award
    - Deployment and expansion to Pilot services
    - Services exposed to end-users and Early Adopters organisations
    - Substantial work on purchasing models: Total Cost Calculators
    - Resulting services available widely to the research community: EOSC exchange

Project Timeline



R&D Competitive Execution

Project Kick-off
Geneva
(07 Feb 2019)

Selected 5 consortia

Selected 3 consortia out of 5

Selected 2 consortia out of 3

Project Ends
(30 June 2022)

Open Market Consultation

R&D Bids Evaluation & Selection

Led by PIC

Design Phase
4 Months

Call Off

Led by EMBL-EBI

Prototype Phase
9 Months

Call Off

Led by DESY

Pilot Phase
7 Months

Pilot Phase Selected Consortia

arkivum
Bringing archived data to life

Google Cloud

libnova

LIBERTAS PERFVNDET
OMNIA LVCE
UNIVERSITAT DE BARCELONA

CSIC
CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

aws

voxility

bidaidea

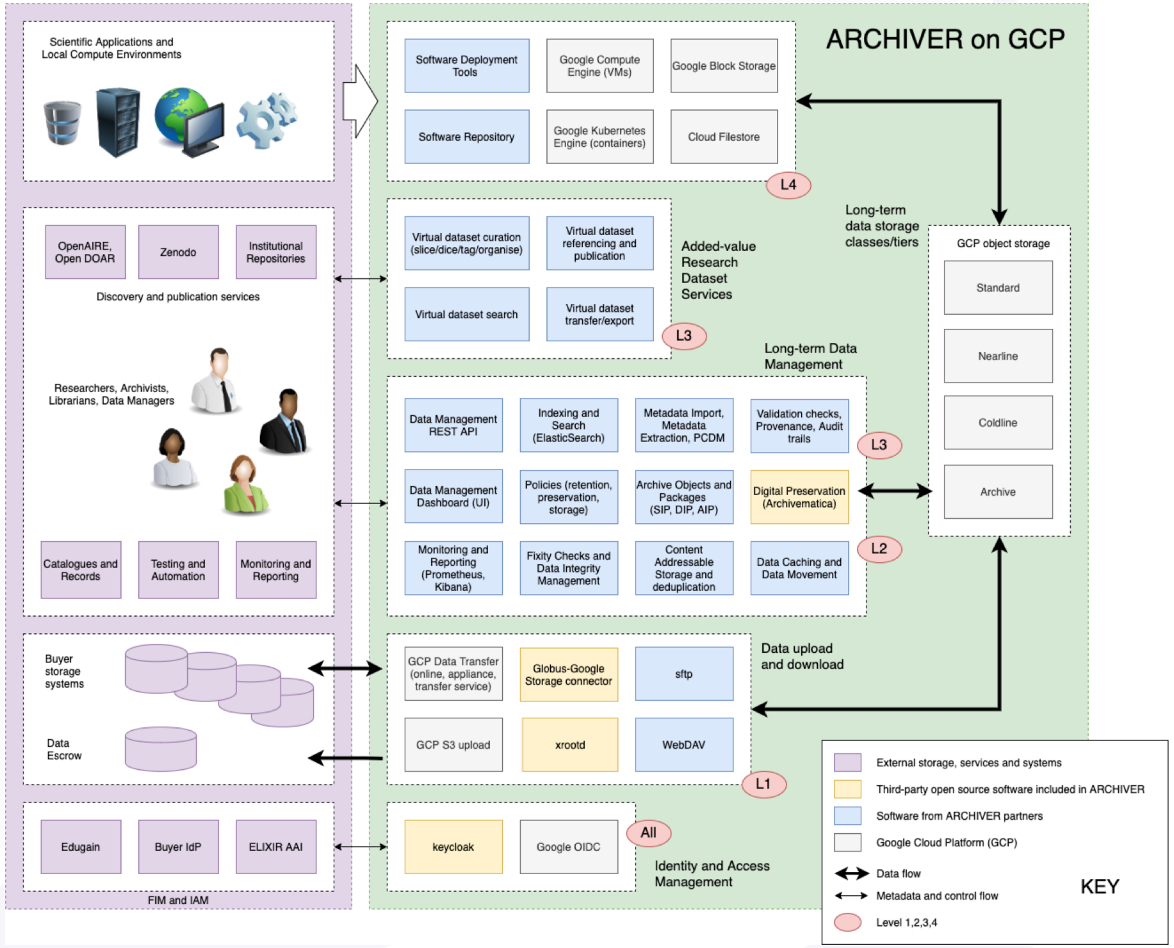
Giaretta Associates

Selected Consortia: **Arkivum**

- Overall architecture composed of micro-services to **scale** to multi-petabyte volumes of billions of objects

- Based on Kubernetes: autoscales, meaning no idle resources which **reduces** costs and carbon emissions

- Different storage options, for example deep archive/cold storage for infrequently accessed data = **cheaper**



Scientific Applications and Local Compute Environments

OpenAIRE, Open DOAR
Zenodo
Institutional Repositories
Discovery and publication services

Researchers, Archivists, Librarians, Data Managers

Catalogues and Records
Testing and Automation
Monitoring and Reporting

Buyer storage systems
Data Escrow

Edugain
Buyer IdP
ELIXIR AAI
FIM and IAM

ARCHIVER on GCP

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

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

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

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

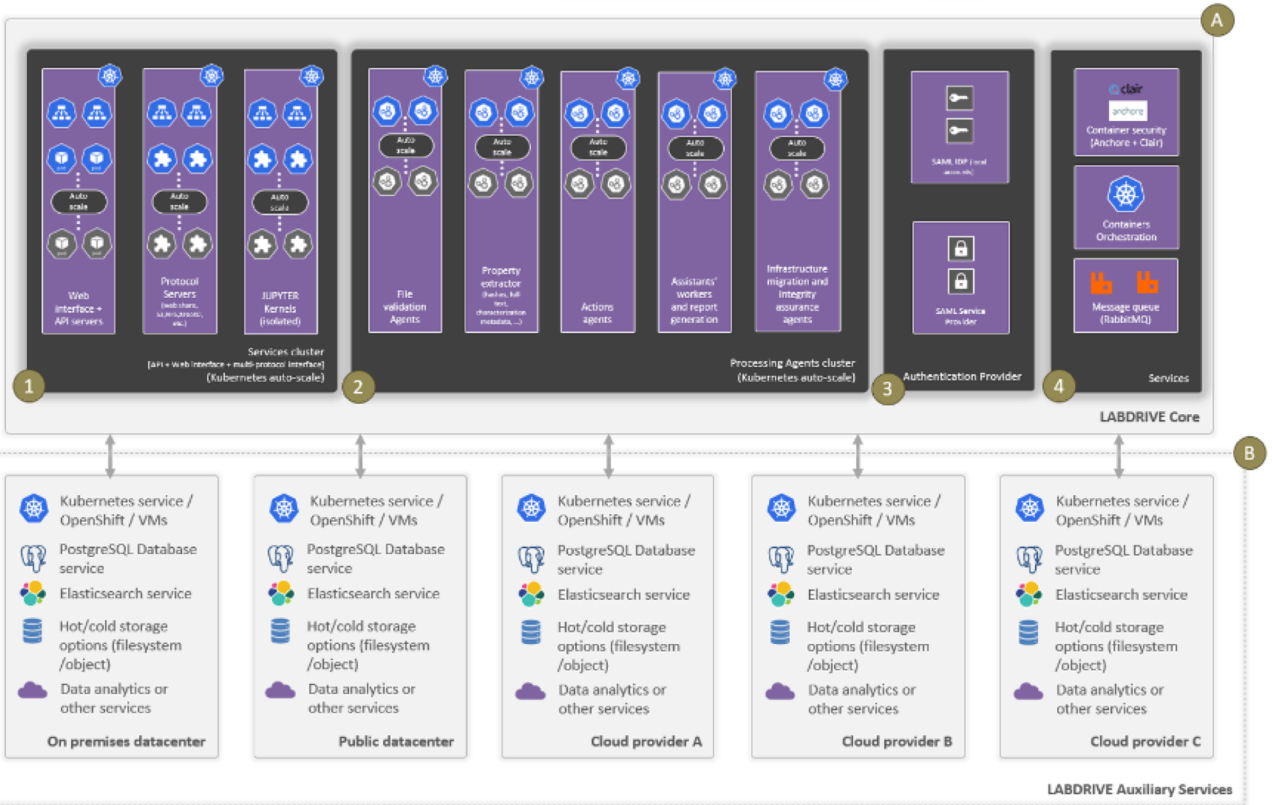
keycloak
Google OIDC
All
Identity and Access Management

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

Prototype architecture of the Arkivum consortium (image courtesy of the Arkivum consortium)

Selected Consortia: **Libnova**

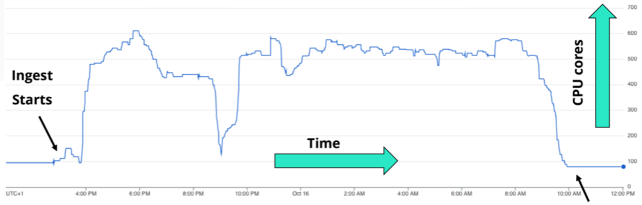
- Prototype based on LibSAFE SaaS
- Using infrastructure provided by AWS; can be deployed on-premises
- Runs on Kubernetes, fully scalable: adjustable number of components based on service demand which translates to cost and environmental effectiveness.
- QoS **optimization** of storage tiers; Less frequently accessed is cheaper


A
Auto scale
Web interface + API servers
Protocol Servers
JUPYTER Kernels (isolated)
Services cluster
[API + Web interface + multi-protocol interface]
(Kubernetes auto-scale)
1
File validation Agents
Property extractor
Actions agents
Assistants' workers and report generation
Infrastructure migration and integrity assurance agents
Processing Agents cluster
(Kubernetes auto-scale)
2
SAML IDP
SAML Service Provider
Authentication Provider
3
clair
anchore
Container security (Anchore + Clair)
Containers Orchestration
Message queue (RabbitMQ)
Services
4
LABDRIVE Core
B
Kubernetes service / OpenShift / VMs
PostgreSQL Database service
Elasticsearch service
Hot/cold storage options (filesystem /object)
Data analytics or other services
On premises datacenter
Public datacenter
Cloud provider A
Cloud provider B
Cloud provider C
LABDRIVE Auxiliary Services

Prototype architecture of the Libnova consortium (image courtesy of the Libnova consortium)

Highlights Pilot Phase (I)

- Scalability & Elasticity
  - Autoscaling of ingest, archiving and preservation: thousands files, 100TB+ per day
  - Maximum profit from elasticity of cloud: efficient matching resources to data processing



- Deployment Models & Security
  - Validation of on-premise/hybrid deployments as a component of exit strategies
  - Security assessments of the resulting services

- FAIR assessments
  - Automated assessments using the F-UJI tool complemented by self-assessments from the consortia

- Science Reproducibility
  - PoC integration with scientific services & apps

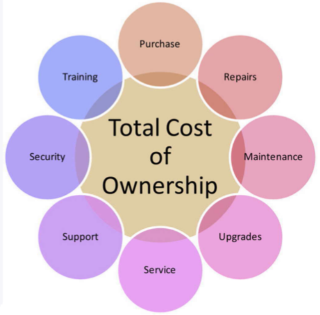
ARCHIVER
ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

Highlights Pilot Phase (II)

- Certification: CoreTrustSeal self-assessments
  - Onboarding feedback received from the External Advisory Board

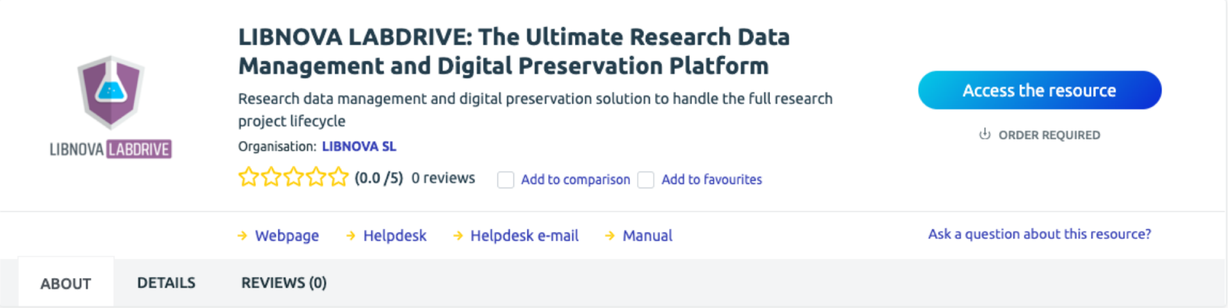
CORE TRUST SEAL

- TCO: Total Cost of Services calculators
  - Allow organisations to estimate the full costs of the solutions for their own use cases
  - Includes carbon footprint analysis



- Availability of resulting R&D to the research community
  - ARCHIVER on EOSC platform Early Adopters Programme
  - ARCHIVER services listed in the EOSC Exchange

Powered by: EOSC Future, C-SCALE, DICE, EGI-ACE, OpenAIRE Nexus, Reliance, ARCHIVER



LIBNOVA LABDRIVE

LIBNOVA LABDRIVE: The Ultimate Research Data Management and Digital Preservation Platform

Research data management and digital preservation solution to handle the full research project lifecycle

Organisation: LIBNOVA SL

☆☆☆☆☆ (0.0 /5) 0 reviews □ Add to comparison □ Add to favourites

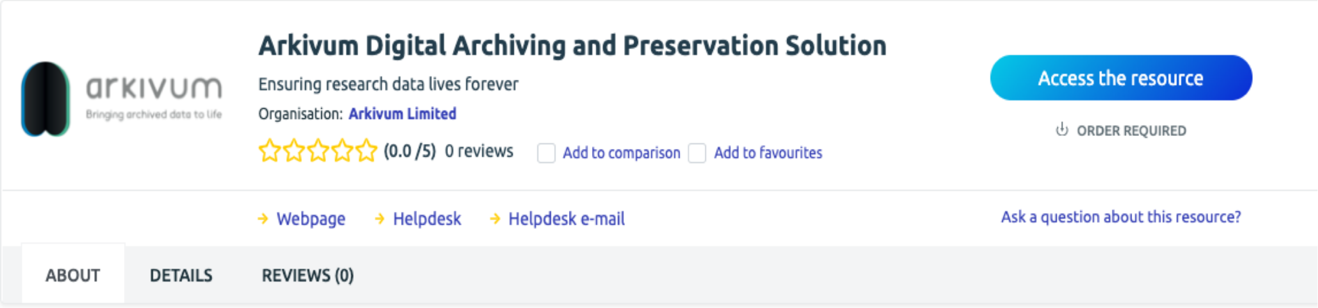
Access the resource

ORDER REQUIRED

→ Webpage → Helpdesk → Helpdesk e-mail → Manual

Ask a question about this resource?

ABOUT DETAILS REVIEWS (0)



arkivum
Bringing archived data to life

Arkivum Digital Archiving and Preservation Solution

Ensuring research data lives forever

Organisation: Arkivum Limited

☆☆☆☆☆ (0.0 /5) 0 reviews □ Add to comparison □ Add to favourites

Access the resource

ORDER REQUIRED

→ Webpage → Helpdesk → Helpdesk e-mail

Ask a question about this resource?

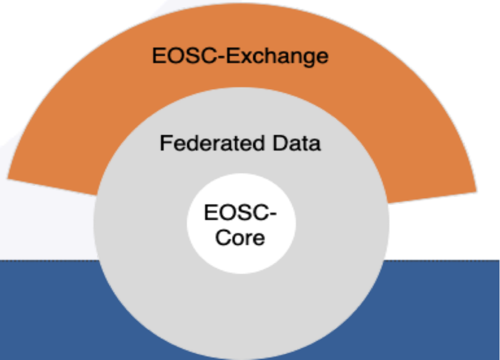
ABOUT DETAILS REVIEWS (0)

Conclusions

- R&D challenge in digital preservation goes beyond data storage
  - Must keep intellectual control of data and make research outputs reusable
  - Must ensure the correct breakdown of responsibilities between Data Stewards and Service Providers
  - Extends FAIR to associated products: Software, Workflows and Infrastructure

- ARCHIVER commoditised LTDP services in the research domain
  - Project objectives successfully met
  - Resulting services and deliverables of contracted companies of high quality

- Sustainable model with services existing beyond ARCHIVER lifetime
  - ARCHIVER services listed in the EOSC Exchange available to the wider community
  - TCO calculators allowing public research organisations to allow for cost planning
  - Progress in adapting cloud purchase models fitting procurement in the public sector



ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

Pilot Phase - Buyers Group use cases

Pilot End Phase

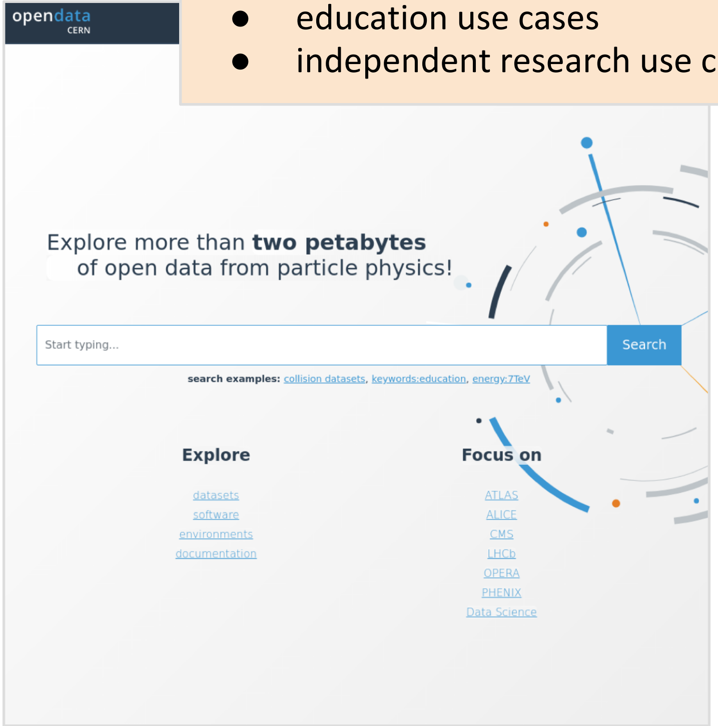
ARCHIVER - Archiving and Preservation for Research Environments project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824516.

CERN - End of Pilot Phase Report

Tibor Simko, Jean-Yves Le Meur, Antonio Vivace, Ignacio Peluaga

CERN Open Data: From data preservation to data reuse

CERN Open Data portal
- more than 2.5 petabytes of particle physics data
- education use cases
- independent research use cases



opendata CERN

Explore more than **two petabytes** of open data from particle physics!

Start typing... Search

search examples: collision datasets, keywords:education, energy:7TeV

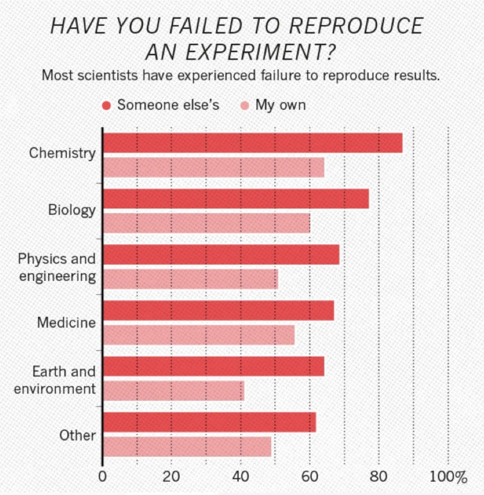
Explore

datasets
software
environments
documentation

Focus on

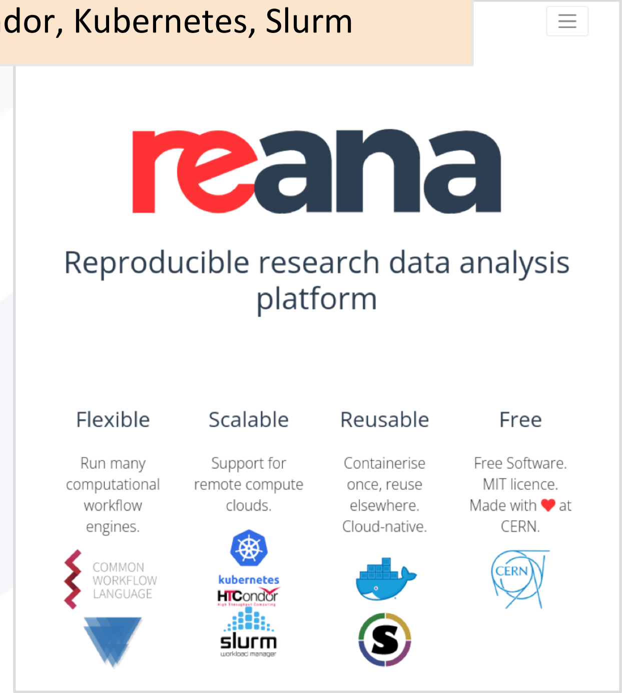
ATLAS
ALICE
CMS
LHCb
OPERA
PHENIX
Data Science

https://opendata.cern.ch



Nature **533** (2016), 452–454
https://doi.org/10.1038/533452a

REANA reproducible analysis platform
- run containerised computational workflows on remote clouds
- CWL, Snakemake, Yadage
- HTCondor, Kubernetes, Slurm



reana

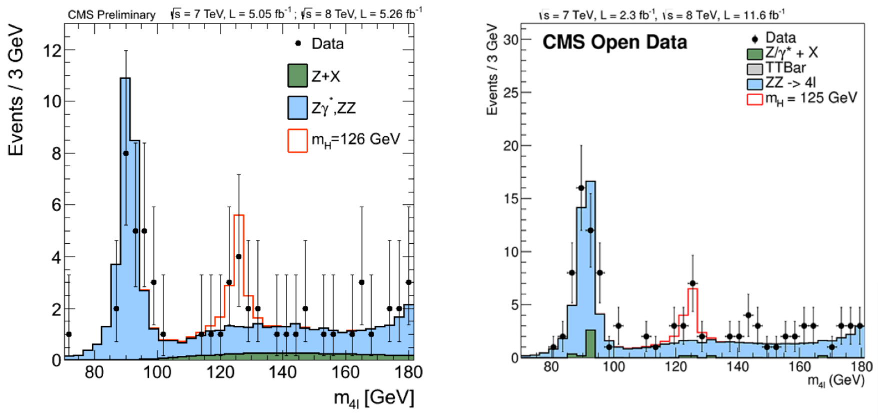
Reproducible research data analysis platform

Flexible	Scalable	Reusable	Free
Run many computational workflow engines.	Support for remote compute clouds.	Containerise once, reuse elsewhere. Cloud-native.	Free Software. MIT licence. Made with ♥ at CERN.

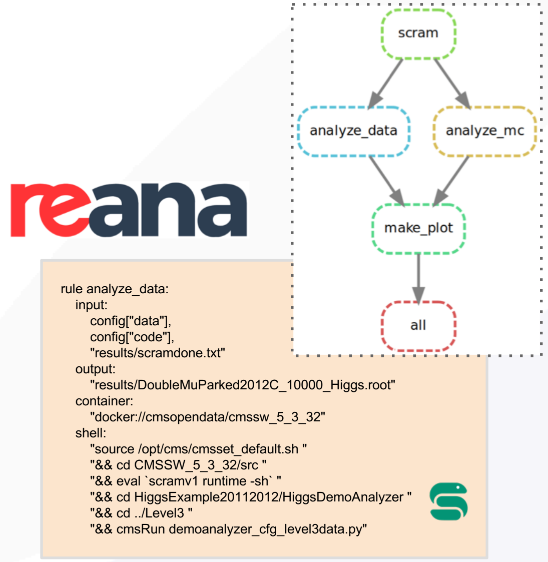
https://www.reana.io

CERN Open Data: From data preservation to data reuse

- testing exposure of (parts of) datasets via XRootD protocol
- testing “Compute” options for content preserved in “Storage”
- reproduced several open data analysis examples
- run Jupyter notebook examples (Python)
- run containerised workflow examples (Docker, Snakemake, Kubernetes)

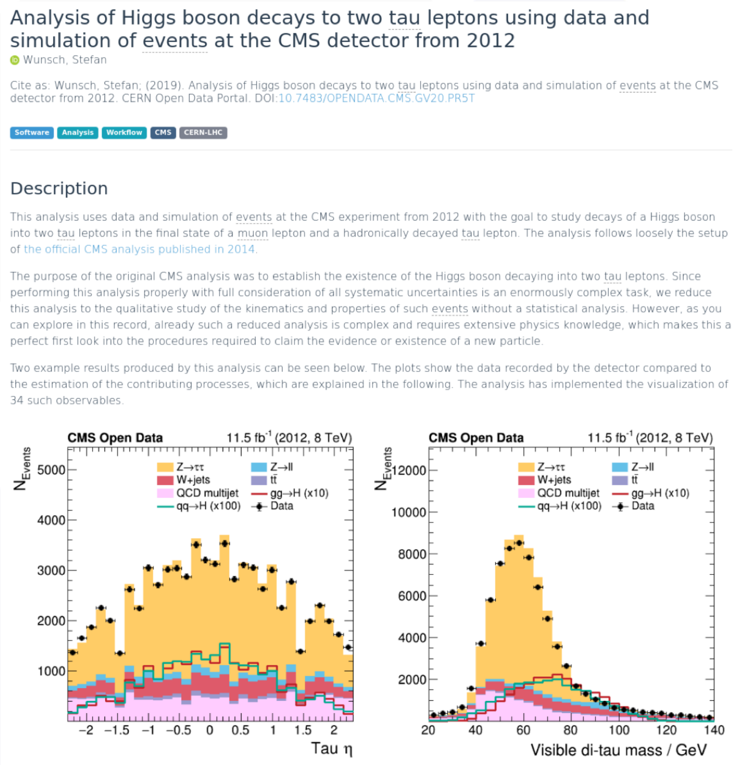


CMS Higgs-to-four-lepton example analysis



```
rule analyze_data:
    input:
        config["data"],
        config["code"],
        "results/scramdone.txt"
    output:
        "results/DoubleMuParked2012C_10000_Higgs.root"
    container:
        "docker://cmsopendata/cmssw_5_3_32"
    shell:
        "source /opt/cms/cmsset_default.sh "
        "&& cd CMSSW_5_3_32/src "
        "&& eval `scramv1 runtime -sh` "
        "&& cd HiggsExample20112012/HiggsDemoAnalyzer "
        "&& cd ../Level3 "
        "&& cmsRun demoanalyzer_cfg_level3data.py"
```

Running containerised workflows



Analysis of Higgs boson decays to two tau leptons using data and simulation of events at the CMS detector from 2012

Wunsch, Stefan

Cite as: Wunsch, Stefan; (2019). Analysis of Higgs boson decays to two tau leptons using data and simulation of events at the CMS detector from 2012. CERN Open Data Portal. DOI:10.7483/OPENDATA.CMS.GV20.PR5T

Software Analysis Workflow CMS CERN-LHC

Description

This analysis uses data and simulation of events at the CMS experiment from 2012 with the goal to study decays of a Higgs boson into two tau leptons in the final state of a muon lepton and a hadronically decayed tau lepton. The analysis follows loosely the setup of the official CMS analysis published in 2014.

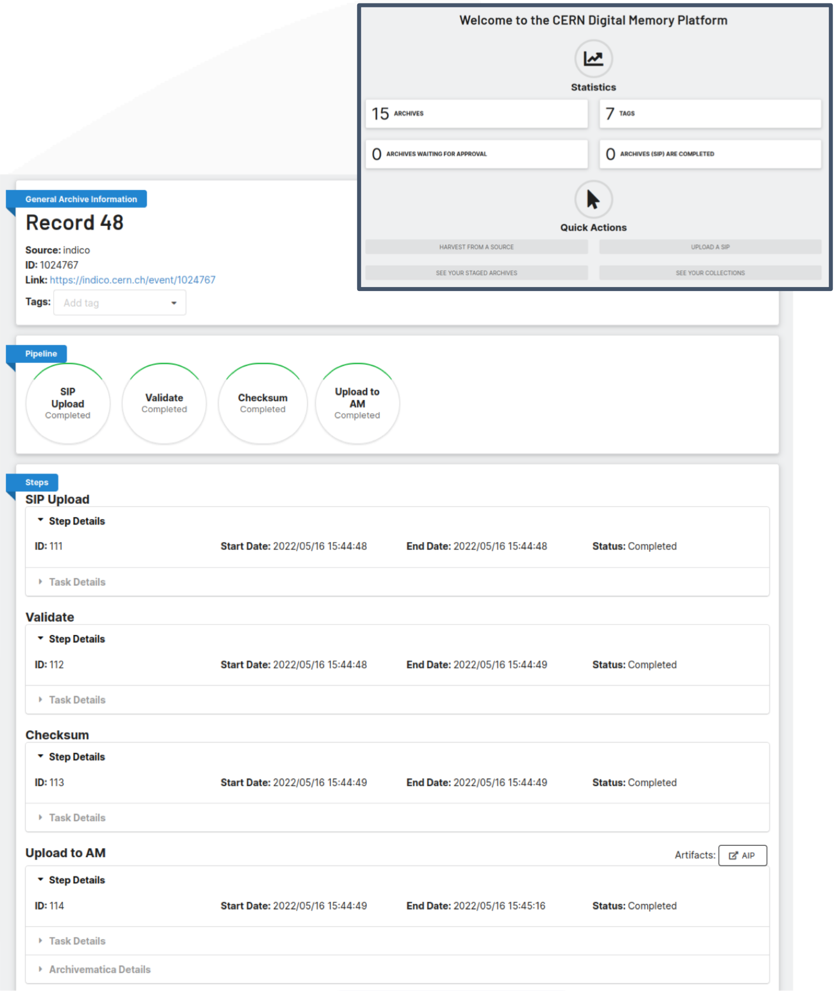
The purpose of the original CMS analysis was to establish the existence of the Higgs boson decaying into two tau leptons. Since performing this analysis properly with full consideration of all systematic uncertainties is an enormously complex task, we reduce this analysis to the qualitative study of the kinematics and properties of such events without a statistical analysis. However, as you can explore in this record, already such a reduced analysis is complex and requires extensive physics knowledge, which makes this a perfect first look into the procedures required to claim the evidence or existence of a new particle.

Two example results produced by this analysis can be seen below. The plots show the data recorded by the detector compared to the estimation of the contributing processes, which are explained in the following. The analysis has implemented the visualization of 34 such observables.

CMS Higgs-to-tautau example analysis

CERN Digital Memory: the OAIS Archive

- **Status**:
  - Data stored in main CERN Information Systems and on local user stores can be harvested to create BagIt-compliant SIPs.
    - Includes publications, preprints, presentations, photos, videos, and more
  - Web Interface to run ''on-demand'' archiving for people leaving CERN
- **Challenges**:
  - Duplication
  - Authorships and Versioning
  - Scaling preservation services/ file conversions to create AIPs



15 ARCHIVES
7 TAGS
0 ARCHIVES WAITING FOR APPROVAL
0 ARCHIVES (SIP) ARE COMPLETED

Record 48

Source: indico
ID: 1024767
Link: https://indico.cern.ch/event/1024767
Tags: Add tag

SIP Upload
ID: 111 **Start Date:** 2022/05/16 15:44:48 **End Date:** 2022/05/16 15:44:48 **Status:** Completed

Validate
ID: 112 **Start Date:** 2022/05/16 15:44:48 **End Date:** 2022/05/16 15:44:49 **Status:** Completed

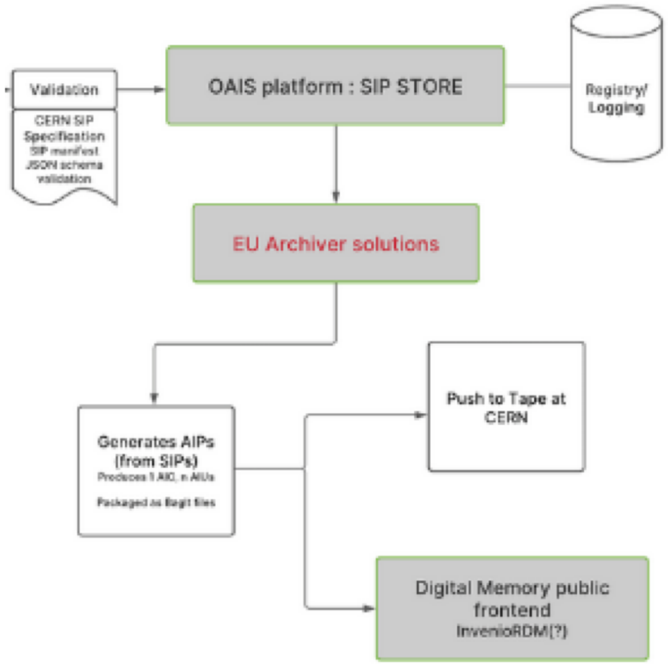
Checksum
ID: 113 **Start Date:** 2022/05/16 15:44:49 **End Date:** 2022/05/16 15:44:49 **Status:** Completed

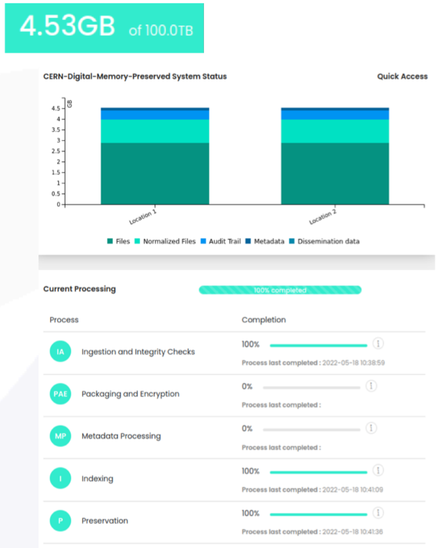
Upload to AM
ID: 114 **Start Date:** 2022/05/16 15:44:49 **End Date:** 2022/05/16 15:45:16 **Status:** Completed

CERN Digital Memory: feeding Archiver.eu

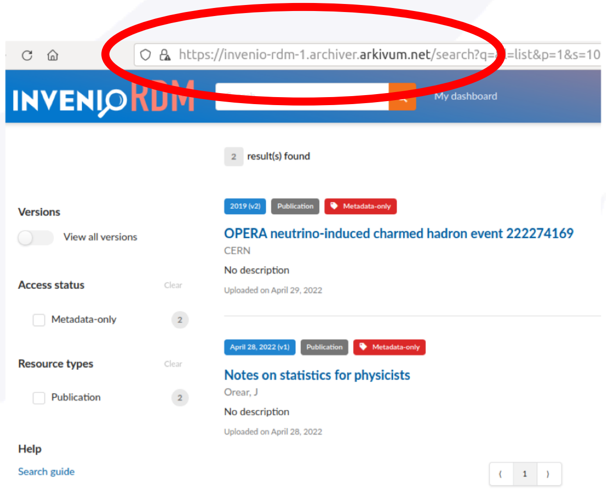
Archiver.eu solutions to ingest CERN SIPs and store preservation artifacts: AIPs, DIPs…

- ➔ Validated metadata manifest technique to transfer main meta information
- ➔ Validated API workflow
- ➔ Validated using Arkivum as an “AIP factory” with Archivematica backend


Validation
CERN SIP Specification SIP manifest JSON schema validation
OAIS platform : SIP STORE
Registry/ Logging
EU Archiver solutions
Generates AIPs (from SIPs)
Produces 1 AIC, n AIUs
Packaged as Bagit files
Push to Tape at CERN
Digital Memory public frontend InvenioRDM(?)



Process	Completion
IA Ingestion and Integrity Checks	100% Process last completed : 2022-05-18 10:38:59
PAE Packaging and Encryption	0% Process last completed :
MP Metadata Processing	0% Process last completed :
I Indexing	100% Process last completed : 2022-05-18 10:41:09
P Preservation	100% Process last completed : 2022-05-18 10:41:36



https://invenio-rdm-1.archiver.arkivum.net/search?q=&l=list&p=1&s=10

2 result(s) found

Versions

View all versions

Access status Clear

□ Metadata-only 2

Resource types Clear

□ Publication 2

Help

Search guide

2019 (v2) Publication Metadata-only

OPERA neutrino-induced charmed hadron event 222274169

CERN

No description

Uploaded on April 29, 2022

April 28, 2022 (v1) Publication Metadata-only

Notes on statistics for physicists

Orear, J

No description

Uploaded on April 28, 2022

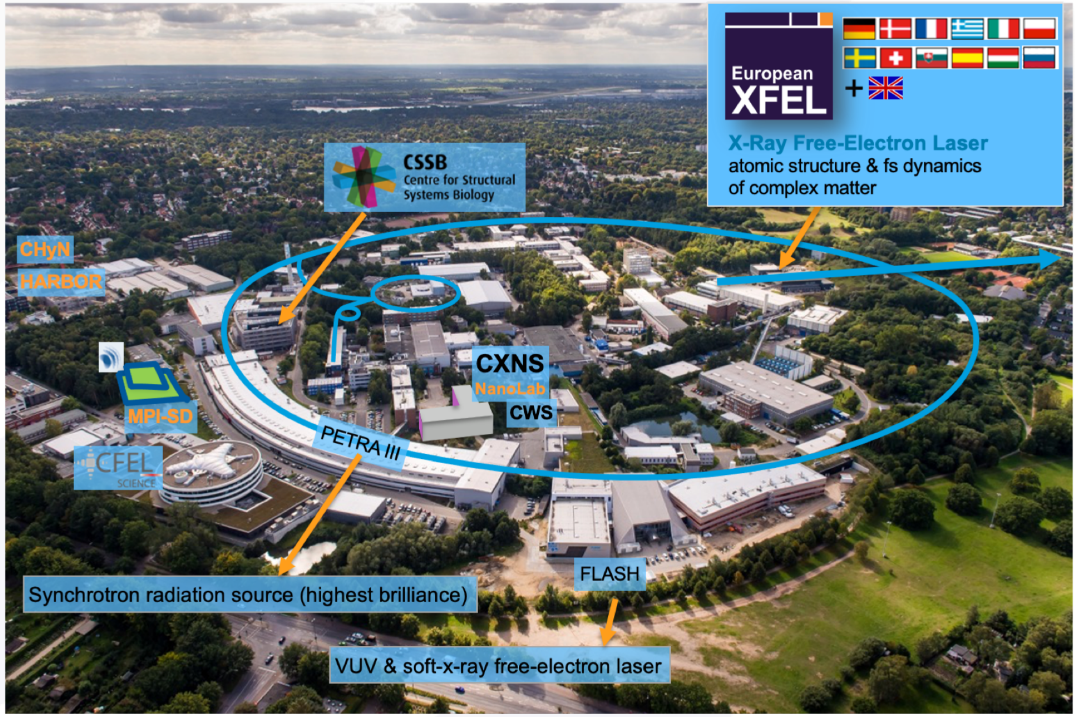
‹ 1 ›

DESY - End of Pilot Phase Report

Martin Gasthuber, Jacek Chodak and Sergey Yakubov

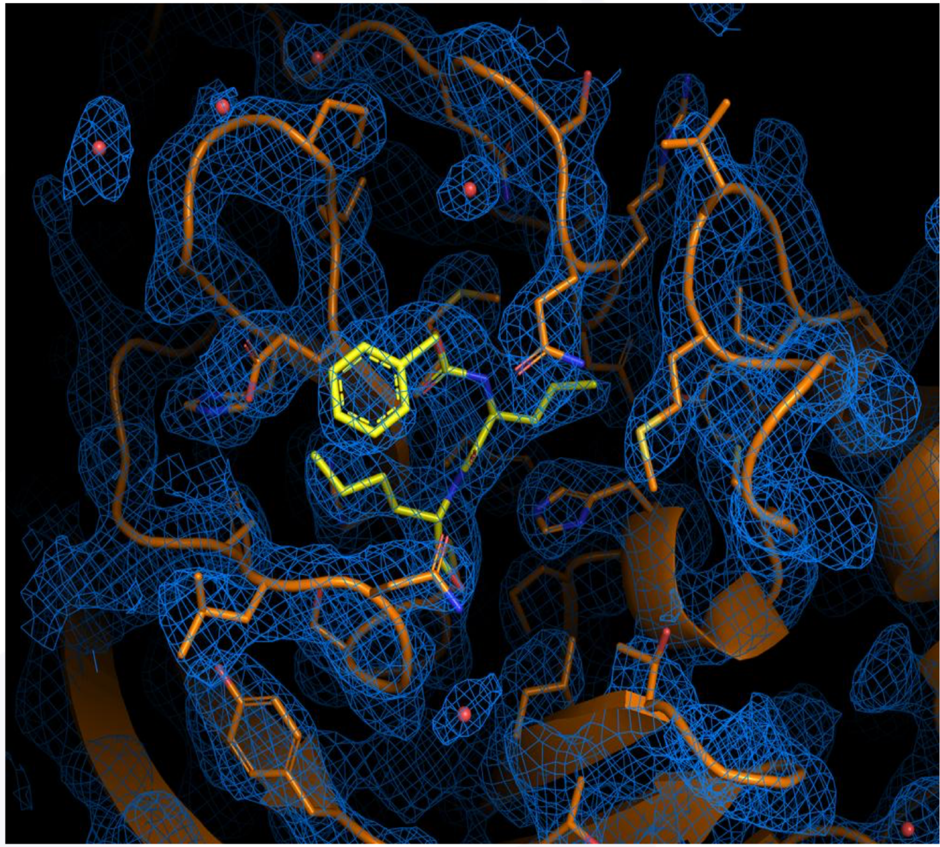
condensed use-cases - all ‘photon science’

- small - individual scientist, interactive
  - PhD thesis, detector developments, …
- mid size ‘set of’ experiments, interactive+API, centrally configured
  - experimental station at synchrotron facility
- large size experiments/facility, API based, centrally configured, automated
  - EuXFEL facility


European XFEL
+
X-Ray Free-Electron Laser
atomic structure & fs dynamics of complex matter
CSSB
Centre for Structural Systems Biology
CHyN
HARBOR
MPI-SD
CFEL
SCIENCE
CXNS
NanoLab
CWS
PETRA III
FLASH
Synchrotron radiation source (highest brilliance)
VUV & soft-x-ray free-electron laser

example of science activities - Corona related

- since early 2020 - shortly after local shutdown started
  - massive X-ray screening, scientists identified promising candidates for drugs against SARS-CoV-2
  - high resolution and three-dimensional representation of damaged lung tissue following severe Covid-19
  - …
- large number of participating scientists from many disciplines requiring 'shared' access, archiving and 'close to' processing power data services and location
  - perfect candidate for ARCHIVER services !



Electron density map of the most antiviral active ingredient calpeptin (yellow) binding at the main protease (Credit: DESY, Sebastian Günther)

pilot phase activities

- scale tests
  - not (yet) at ‘facility’ level - ~10 PiB per week demand today
    - sadly enough, accelerated demand rate since 2019
    - tests at that scale are still prohibitive today (>16GB/s contiguous network BW)
  - other two fits well in capacity / #of files, challenging (10M files per exp.)
- API - for mid and large with automated access
- templated and hierarchical metadata structures with ‘mandatory’ and ‘optional’ elements allowing user extensions
- on-site deployments - base for hybrid configuration
  - i.e. for second copy in public cloud -> easy and efficient handling of ‘open data’

EMBL-EBI - End of Pilot Phase Report

Justin Clark-Casey

Use case pursued in the Pilot Phase

EMBL on FIRE

- Original concept (use Archiver as an additional FIRE replica) invalidated by FIRE design changes
- Instead pilot using Archiver as a different type of storage for FIRE
  - Archival (there but not immediately accessible) instead of immediately online
  - Put control into the hands of the (internal) users
  - Arguably closer to the core Archiver concept

Pilot Phase testing

EMBL on FIRE

- Internal teams make the decision on when to archive data
  - To reduce use of their main storage/FIRE quotas
- They operate the process (CLI) but must submit a schema containing mandatory elements
  - e.g. Dublin Core identifier, publisher
- Functionality works – need to do more exposure to teams

Pilot Phase testing

EMBL on FIRE

- Teams can request export of data back to main storage
- Permissioned: teams request and administrator approves
- Operates through the Arkivum UI
- Functionality works – needs more exposure to teams

Pilot Phase testing

EMBL on FIRE

- From the prototype phase we know that the contractors can handle data at scale
- And we have knowledge of the costs involved
- We need to continue piloting with internal users
- And fit this into a fast evolving internal storage architecture

PIC- End of Pilot Phase Report

Jordi Casals, Manuel Delfino, Meritxell Garcia

Use cases (from realistic needs of MAGIC telescopes)[1,2]

4.1 Safe-keeping and recall of large-size files

- Use the platform for daily safe-keeping of scientific data
- Ability to define metadata schemas on upload
- Transfer 1TB-2TB per day with throughputs above 1Gbps/day
- Perform metadata based searches
- Bulk download must be available
- Metadata based downloads of filesets

4.2 Archiving, preservation and distribution of datasets

Same requirements as 4.1 plus:

- Unpack and process *tar* files containing *bagit* bags
- Assign different metadata globally and file by file
- User belong to different/multiple groups
- Select subsets of files based on metadata and share them with other users/groups
- Subsets may be open to be shared publicly

4.3 In-archive processing

- Create automatic processes to be performed on upload
- Run processes over the uploaded data on demand
- Different tools as Jupyter Notebooks, batch processing, etc
- Ability to do all that without having to download, process and reupload

Tests performed

4.1 Safe-keeping and recall of large-size files

- Automatic daily uploads simulating uploads from the telescope
  - File by file every night (simulated flow from telescope data acquisition)
  - Groups of files (datasets) using *bagit* bags (simulated flow from data center)
- Metadata association tested in two different manners
  - Associate metadata *a posteriori* to individual uploaded files
  - Process metadata on *bagit* bag upload (metadata file as part of the bag)
- Tests of rapidly downloading entire datasets to fulfill two scenarios
  - (PIC) Data center disaster recovery
  - Provide input for data reprocessing (off-premise becomes primary archive)

Tests performed

4.2 Archiving, preservation and distribution of datasets

- Common features with Use Case 4.1:
  - *bagit* bags upload method including the file metadata
  - Assigning/updating metadata *a posteriori* to uploaded data
- Reduced data → smaller individual files
  → larger number of files per dataset → using *tar*-ed bags becomes essential
- Web Portals allow browsing and casual searches of archive
  - Portal file listings have direct links for downloading individual files
- Complex metadata searches define data subsets which can be downloaded
  - Enable power users to make complex metadata queries via command-line
  - Simplify (and make reliable) dataset downloading using *bagit*, *tar*/*zip*, downloading scripts
- Open datasets enable distribution to anonymous users

Tests performed

4.3 In-archive processing

- Move data processing from PIC to the Cloud hosting the archive (would avoid downloading data for reprocessing in Use Case 4.1)
- Processing can be triggered using cloud storage events on upload
- Jupyter Notebooks hosted in the archive Cloud with direct access to the archived data has enormous potential:
  - Provide elastic analysis resources to scientific collaborators
  - Flexible access control in archive enables selective sharing between different groups of researchers
  - Provide data and software together for Open Data and dissemination
- Cost and carbon calculators will help guide purchasing decisions

Future plans and conclusions

- We have been able to do everything we expected, though not necessarily the way we originally thought.
- We learned a lot from the contractors, and we think that they learned a lot from us as well.
- We required *bagit* format rolled into *tar* files to upload datasets
  - In OAIS terminology, this is our SIP format
  - ~100x less files to upload eases monitoring and error correction
- We ended up using a simple ad-hoc metadata schema
  - Metadata standards still emerging, current offerings seem too complex for our cases
  - MAGIC bags contain file-by-file metadata in a particular *json* format. Contractors developed their systems to handle this
  - Exported data subsets containing file-by-file metadata boost useability
- Both systems have different but powerful approaches
  - Potential for excellent final products for archiving and preservation of big-science data

PIC
port d'informació
científica

Helping to turn Information into Knowledge

Please visit us at https://www.pic.es

PIC is maintained by a collaboration agreement between Institut de Física d'Altes Energies (IFAE) and Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT) with additional support from Universitat Autònoma de Barcelona (UAB).

ARCHIVER

ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

Early Adopter Use Case: ECMWF Open Data

Raoult Baudouin (ECMWF)

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.

BREAK

ARCHIVER

ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

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

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.

libnova

the most advanced digital preservation platform

ARCHIVER

ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

PILOT END PHASE EVENT

2022-06-13

Big thanks to the Archiver team, EU Representatives and the LIBNOVA Consortium team.

How are you and the ARCHIVER project benefiting LIBNOVA?

Market

Time to market:

Shorter development lifecycle leading to a faster time to market (5 to 2 years). Ahead of every other competitors.

Tangible results:

Without the Archiver project our product would not have been ready until 2024.

Instead, **3 USA/EU top Universities and BAYER** (pharmaceutical) already purchased it, and we are in negotiations with other 12 customers.

Technology

Accelerated innovation:

Being able to work with the 4 buyers (use cases) has maximized our understanding of our customer needs.

Tangible results:

Focused on providing the most advanced digital preservation platform and the best possible service. We have been pushing the boundaries of what is possible in digital preservation, incorporating innovations that empower the organizations to preserve their content in an easier and more efficient way.

Team

Strong partnership:

By working together in the consortium, we are creating strong relationships and incremental business.

Tangible results:

Partnership agreement with Amazon AWS and presence in EOSC marketplace. Strong business perspectives on working together. Future work together with David Giaretta, UB, CSIC, Bidaidea and Voxility.

Everyone working really hard, and really happy about it!

What is the result of the performed R&D?

Scalable platform

15PB

Successful
ingestion test

The platform has been tested to ingest 15 petabytes of content in 30 days, at the remarkable performance of 500TB/day

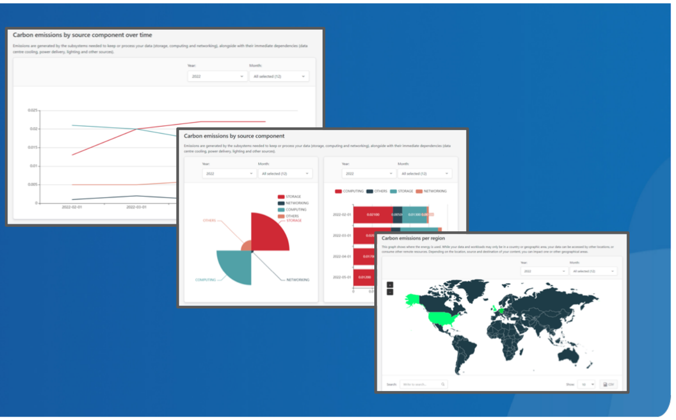
Full reproducibility capabilities

LABDRIVE allow organizations to keep their CONTENT and they CODE in the same platform, with the successful integration with CERN's Reana (and Snakemake) and the capability of running Jupyter notebooks in the platform.

Environmentally friendly

Lowest possible environmental impact when not in use.

Platform carbon footprint evaluated in **real time.**


Carbon emissions by source component over time

Carbon emissions by source component

Carbon emissions per region

Based on standards

First provider in the market to achieve ISO27001, ISO27017, ISO 27018 (all of them achieved during Archiver project).
In the process of ISO 16363 certification.

ISO
16363

9.001

27.001

27.017

27.018

Commercialization plan

Actions

- LABDRIVE is a resource already available in the EOSC Marketplace
- LABDRIVE is a resource listed in AWS Marketplace
- Demo platform available to get “your hands dirty”
- Enlarge the current installed (or ready to install) base
- A marketing/communication campaign is planned, but with no dates yet

Goals and activities to perform

- Raise awareness of the importance of preserving research datasets during their entire life cycle
- Launch a campaign to attract new users to LABDRIVE in EU and US
- Position LIBNOVA as the leading provider of Research Data Management and Digital Preservation solutions
- Increase number of current partnerships / alliances

libnova
the most advanced digital preservation platform

Thanks again!

contact@libnova.com

ARCHIVER

HELIX NEBULA

ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

Arkivum - Google Cloud

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.

arkivum

Bringing archived data to life

The Arkivum solution for ARCHIVER

Matthew Addis, Tom Lynam
Arkivum

Agenda

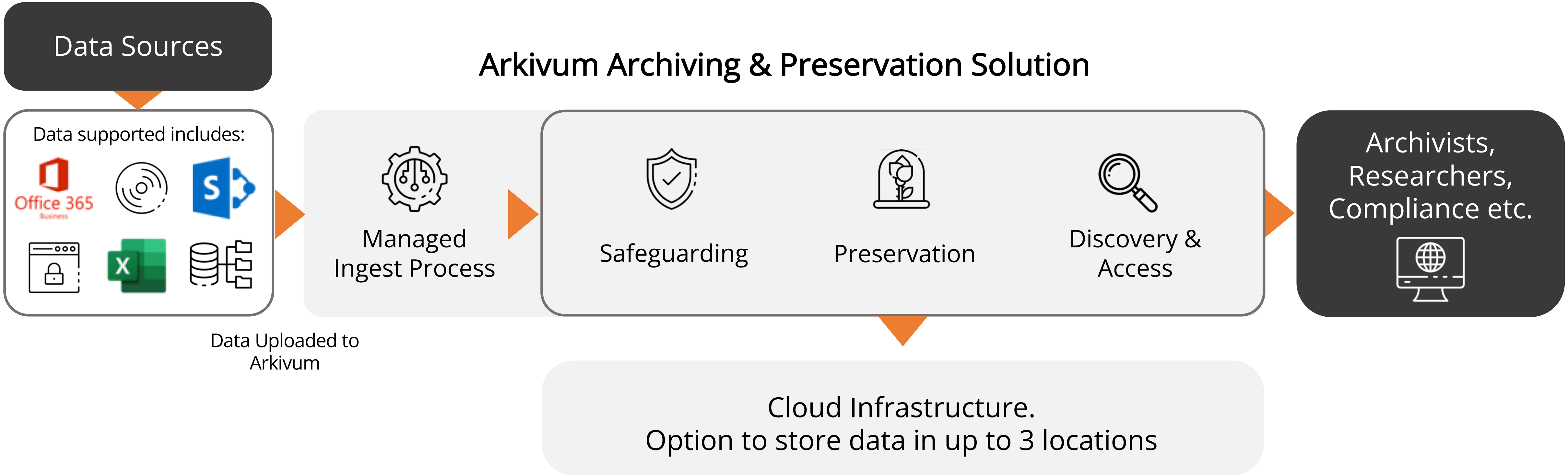
- Overview of Arkivum
- The Arkivum solution for ARCHIVER
- Commercialisation plans
- Questions and discussion

Arkivum and our Product and Services for Long Term Digital Preservation

About Arkivum

- Founded in 2011 out of the University of Southampton – initial focus on Higher Education Research Data
- VC backed and funded
- Headquartered in Reading, UK
- ~25 full time employees
- ISO 9001 and 27001 certified
- Full SaaS offering launched in 2017
- +50 customers as of 2022 across Higher Education, Heritage, Pharma and Life Sciences

Arkivum Fully Managed Archiving and Preservation Solution



Data Sources

Data supported includes:

Data Uploaded to Arkivum

Arkivum Archiving & Preservation Solution

Managed Ingest Process

Safeguarding

Preservation

Discovery & Access

Archivists, Researchers, Compliance etc.

Cloud Infrastructure.
Option to store data in up to 3 locations

Standards, Certification, Quality and Assessment

- Certified to ISO 9001 and ISO 27001
- SoPs for everything we do as part of our QMS
- Validated product releases
- GxP audited by customers, with a 100% success rate
- Mappings available of our product/services to NDSA preservation levels, DPC RAM, Core Trust Seal and FAIR

CfA Centre for Assessment ISO 9001 19/0716 UKAS MANAGEMENT SYSTEMS 0120

CfA Centre for Assessment ISO 27001 17/1349 UKAS MANAGEMENT SYSTEMS 0120

NDSA

Levels of Digital Preservation

dpc

CORE TRUST SEAL

FAIRSFAIR

Fostering Fair Data Practices in Europe

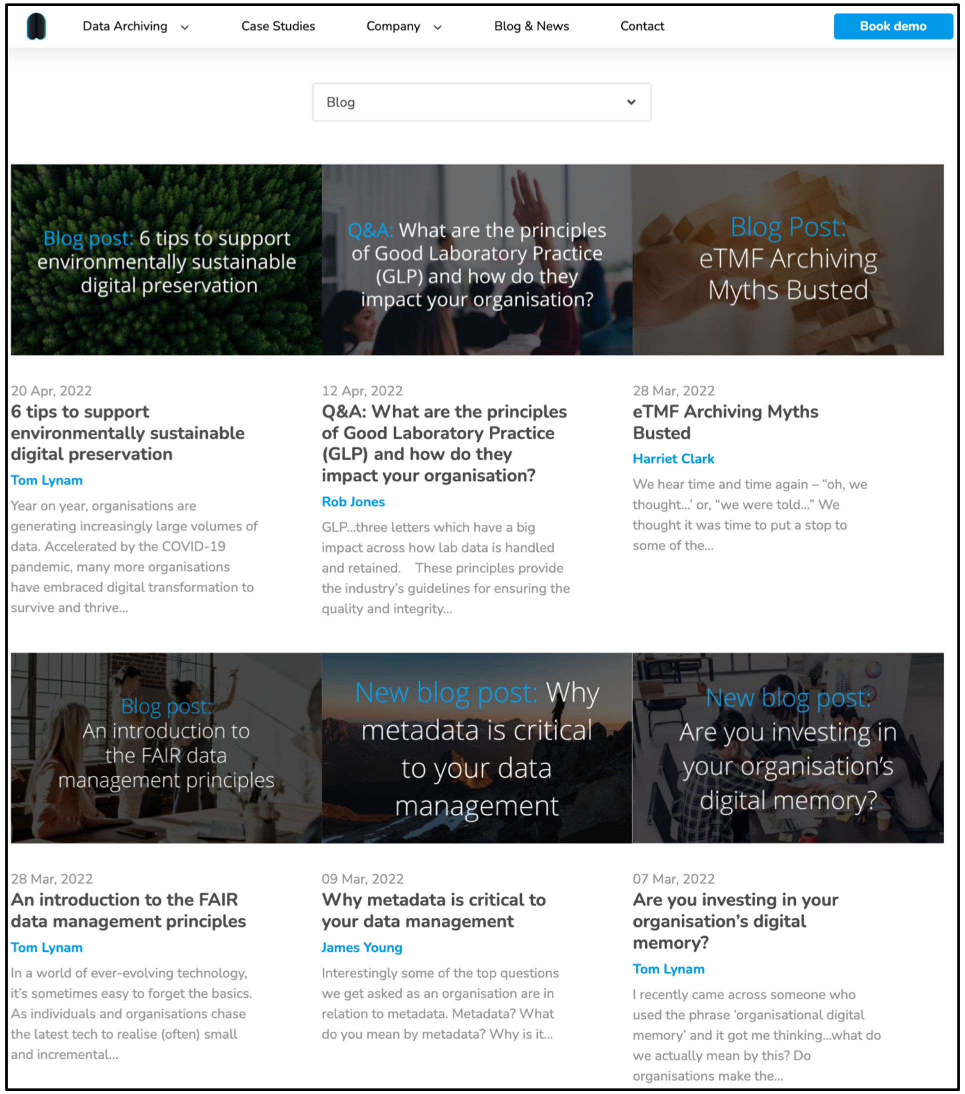
Services and Good Practice

Professional and Customer Services

- Requirements analysis, SoPs and workflows
- Onboarding
- Data migrations
- Computer Systems Validation
- Integrations

Advocacy, Education, Advice

- Regular blog content
- eBooks
- Monthly webinars
- Research reports
- Conferences and events
- Promotion of LTDP practices into in new domains



Blog post: 6 tips to support environmentally sustainable digital preservation

Q&A: What are the principles of Good Laboratory Practice (GLP) and how do they impact your organisation?

Blog Post: eTMF Archiving Myths Busted

20 Apr, 2022
6 tips to support environmentally sustainable digital preservation
Tom Lynam
Year on year, organisations are generating increasingly large volumes of data. Accelerated by the COVID-19 pandemic, many more organisations have embraced digital transformation to survive and thrive...

12 Apr, 2022
Q&A: What are the principles of Good Laboratory Practice (GLP) and how do they impact your organisation?
Rob Jones
GLP...three letters which have a big impact across how lab data is handled and retained. These principles provide the industry's guidelines for ensuring the quality and integrity...

28 Mar, 2022
eTMF Archiving Myths Busted
Harriet Clark
We hear time and time again – "oh, we thought...' or, "we were told..." We thought it was time to put a stop to some of the...

Blog post: An introduction to the FAIR data management principles

New blog post: Why metadata is critical to your data management

New blog post: Are you investing in your organisation's digital memory?

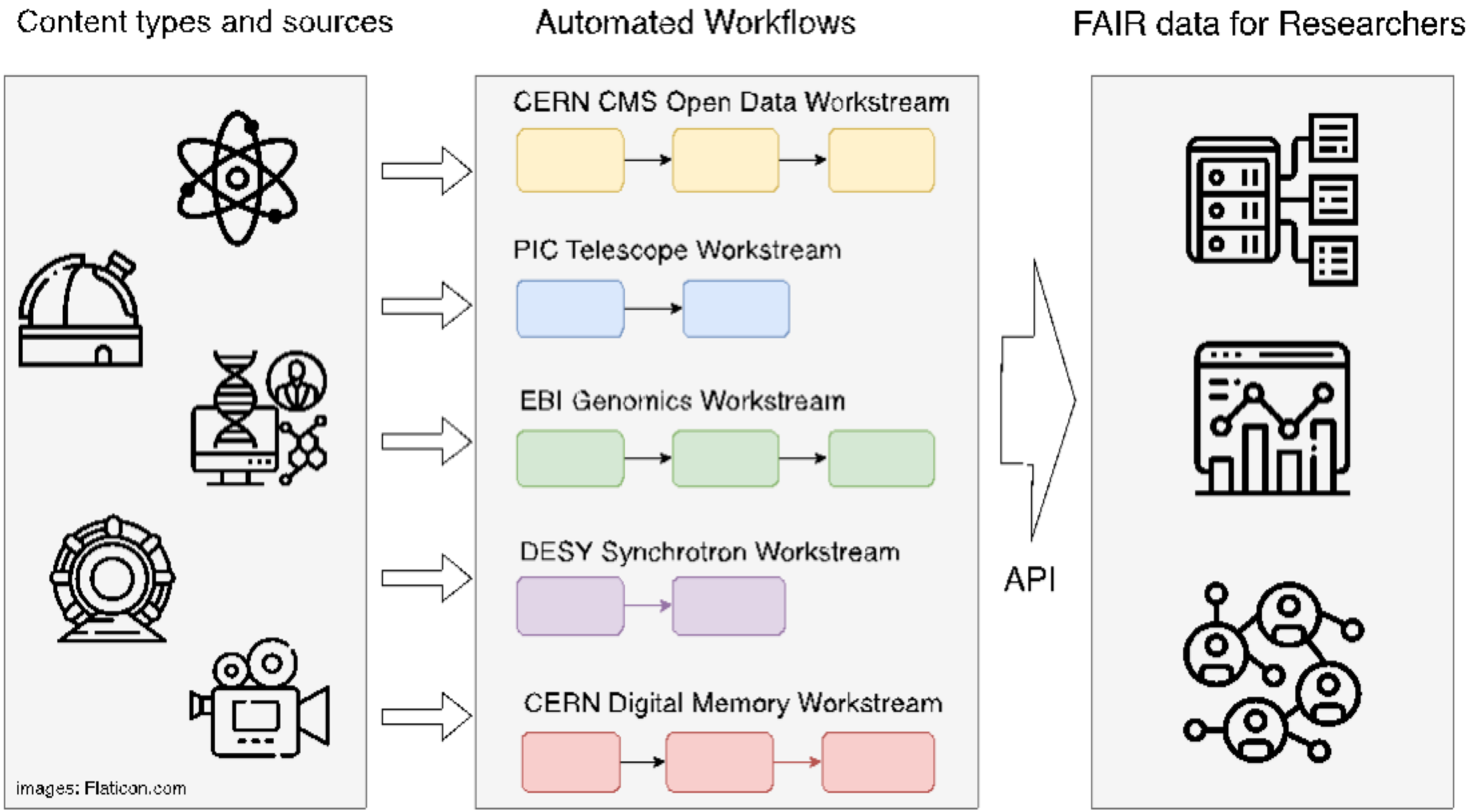
28 Mar, 2022
An introduction to the FAIR data management principles
Tom Lynam
In a world of ever-evolving technology, it's sometimes easy to forget the basics. As individuals and organisations chase the latest tech to realise (often) small and incremental...

09 Mar, 2022
Why metadata is critical to your data management
James Young
Interestingly some of the top questions we get asked as an organisation are in relation to metadata. Metadata? What do you mean by metadata? Why is it...

07 Mar, 2022
Are you investing in your organisation's digital memory?
Tom Lynam
I recently came across someone who used the phrase 'organisational digital memory' and it got me thinking...what do we actually mean by this? Do organisations make the...

Arkivum Solution for ARCHIVER

Long-Term Digital Preservation (LTDP) Factories



Content types and sources

Automated Workflows

FAIR data for Researchers

CERN CMS Open Data Workstream

PIC Telescope Workstream

EBI Genomics Workstream

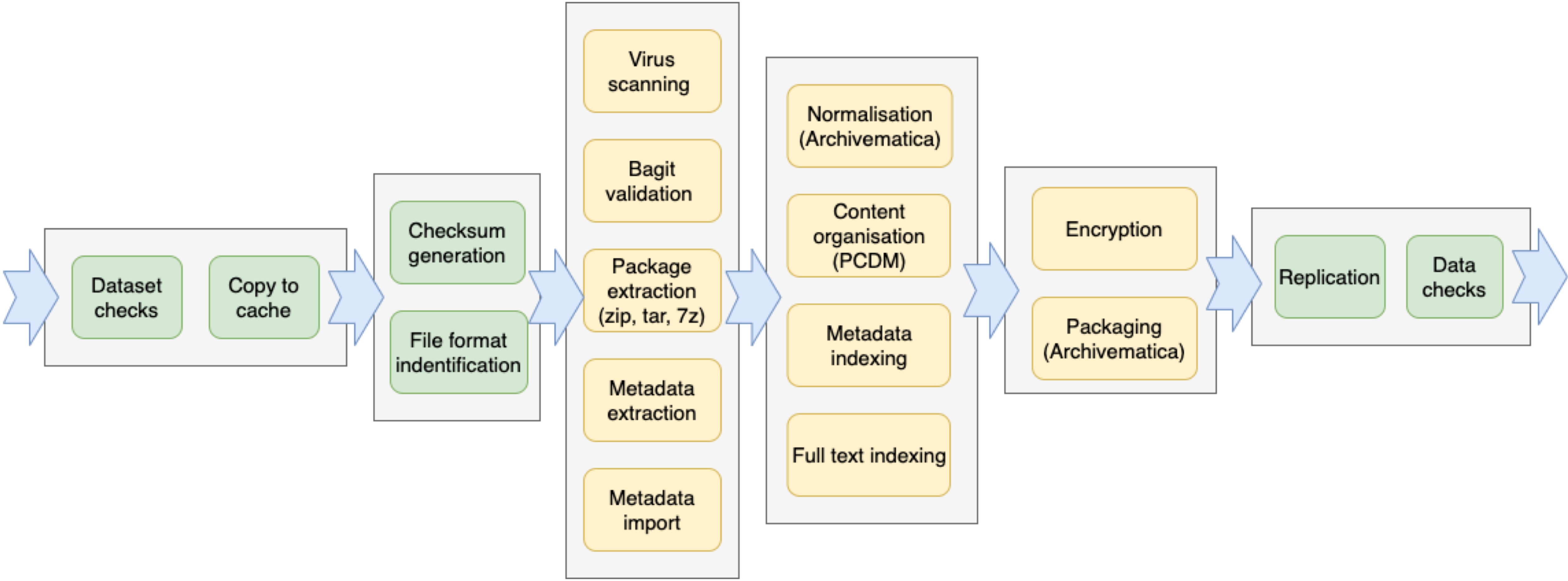
DESY Synchrotron Workstream

CERN Digital Memory Workstream

API

images: Flaticon.com

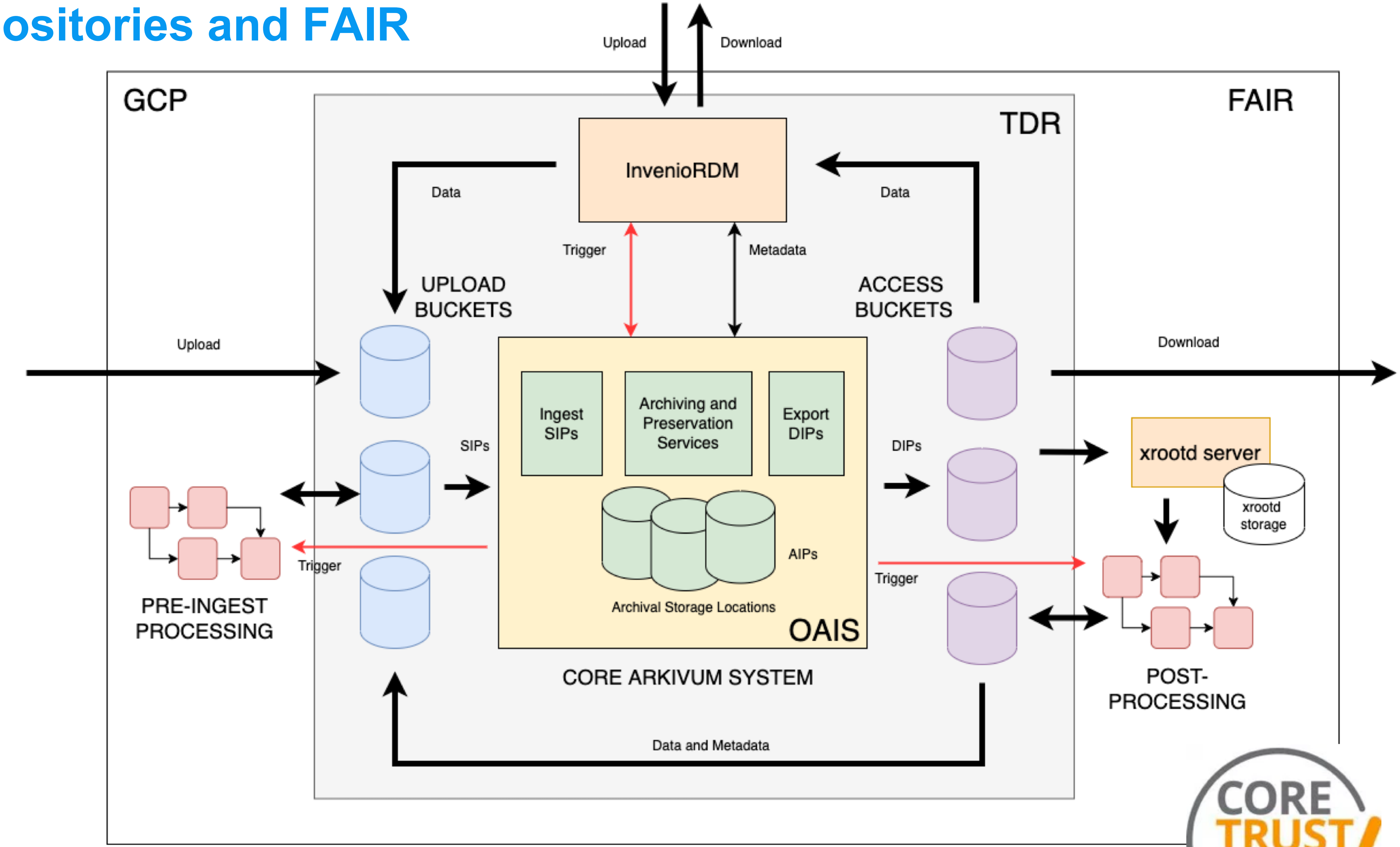
Arkivum Microservices and Workflows

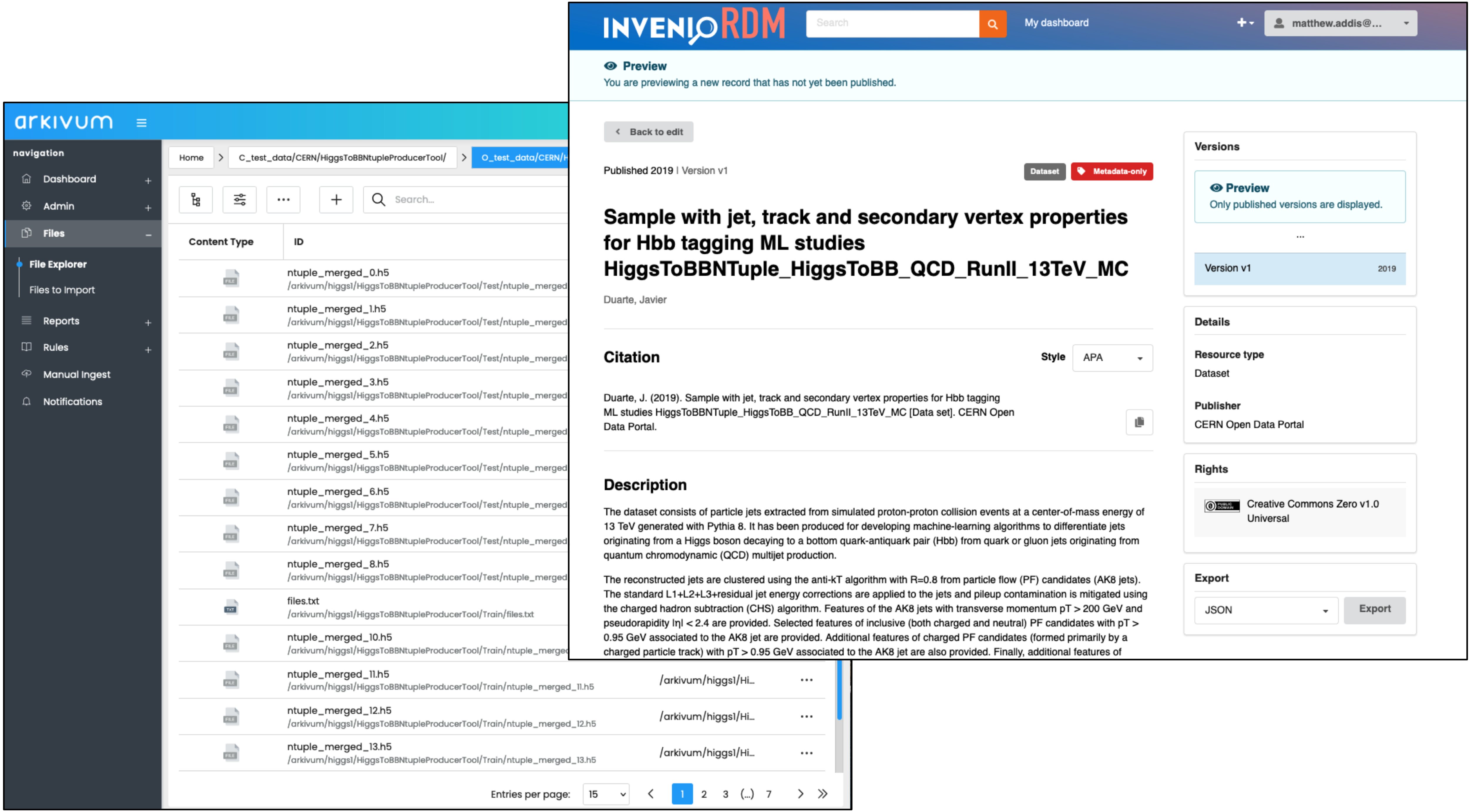


NDSA

Levels of Digital Preservation

Repositories and FAIR





Content Type	ID		
FILE	ntuple_merged_0.h5 /arkivum/higgs1/HiggsToBBNtupleProducerTool/Test/ntuple_merged		
FILE	ntuple_merged_1.h5 /arkivum/higgs1/HiggsToBBNtupleProducerTool/Test/ntuple_merged		
FILE	ntuple_merged_2.h5 /arkivum/higgs1/HiggsToBBNtupleProducerTool/Test/ntuple_merged		
FILE	ntuple_merged_3.h5 /arkivum/higgs1/HiggsToBBNtupleProducerTool/Test/ntuple_merged		
FILE	ntuple_merged_4.h5 /arkivum/higgs1/HiggsToBBNtupleProducerTool/Test/ntuple_merged		
FILE	ntuple_merged_5.h5 /arkivum/higgs1/HiggsToBBNtupleProducerTool/Test/ntuple_merged		
FILE	ntuple_merged_6.h5 /arkivum/higgs1/HiggsToBBNtupleProducerTool/Test/ntuple_merged		
FILE	ntuple_merged_7.h5 /arkivum/higgs1/HiggsToBBNtupleProducerTool/Test/ntuple_merged		
FILE	ntuple_merged_8.h5 /arkivum/higgs1/HiggsToBBNtupleProducerTool/Test/ntuple_merged		
TXT	files.txt /arkivum/higgs1/HiggsToBBNtupleProducerTool/Train/files.txt		
FILE	ntuple_merged_10.h5 /arkivum/higgs1/HiggsToBBNtupleProducerTool/Train/ntuple_merge		
FILE	ntuple_merged_11.h5 /arkivum/higgs1/HiggsToBBNtupleProducerTool/Train/ntuple_merged_11.h5	/arkivum/higgs1/Hi...	...
FILE	ntuple_merged_12.h5 /arkivum/higgs1/HiggsToBBNtupleProducerTool/Train/ntuple_merged_12.h5	/arkivum/higgs1/Hi...	...
FILE	ntuple_merged_13.h5 /arkivum/higgs1/HiggsToBBNtupleProducerTool/Train/ntuple_merged_13.h5	/arkivum/higgs1/Hi...	...

Preview

You are previewing a new record that has not yet been published.

Published 2019 | Version v1

Sample with jet, track and secondary vertex properties for Hbb tagging ML studies HiggsToBBNTuple_HiggsToBB_QCD_RunII_13TeV_MC

Duarte, Javier

Citation

Style: APA

Duarte, J. (2019). Sample with jet, track and secondary vertex properties for Hbb tagging ML studies HiggsToBBNTuple_HiggsToBB_QCD_RunII_13TeV_MC [Data set]. CERN Open Data Portal.

Description

The dataset consists of particle jets extracted from simulated proton-proton collision events at a center-of-mass energy of 13 TeV generated with Pythia 8. It has been produced for developing machine-learning algorithms to differentiate jets originating from a Higgs boson decaying to a bottom quark-antiquark pair (Hbb) from quark or gluon jets originating from quantum chromodynamic (QCD) multijet production.

The reconstructed jets are clustered using the anti-kT algorithm with R=0.8 from particle flow (PF) candidates (AK8 jets). The standard L1+L2+L3+residual jet energy corrections are applied to the jets and pileup contamination is mitigated using the charged hadron subtraction (CHS) algorithm. Features of the AK8 jets with transverse momentum pT > 200 GeV and pseudorapidity |η| < 2.4 are provided. Selected features of inclusive (both charged and neutral) PF candidates with pT > 0.95 GeV associated to the AK8 jet are provided. Additional features of charged PF candidates (formed primarily by a charged particle track) with pT > 0.95 GeV associated to the AK8 jet are also provided. Finally, additional features of

Versions

Preview

Only published versions are displayed.

...

Version v1 2019

Details

Resource type: Dataset

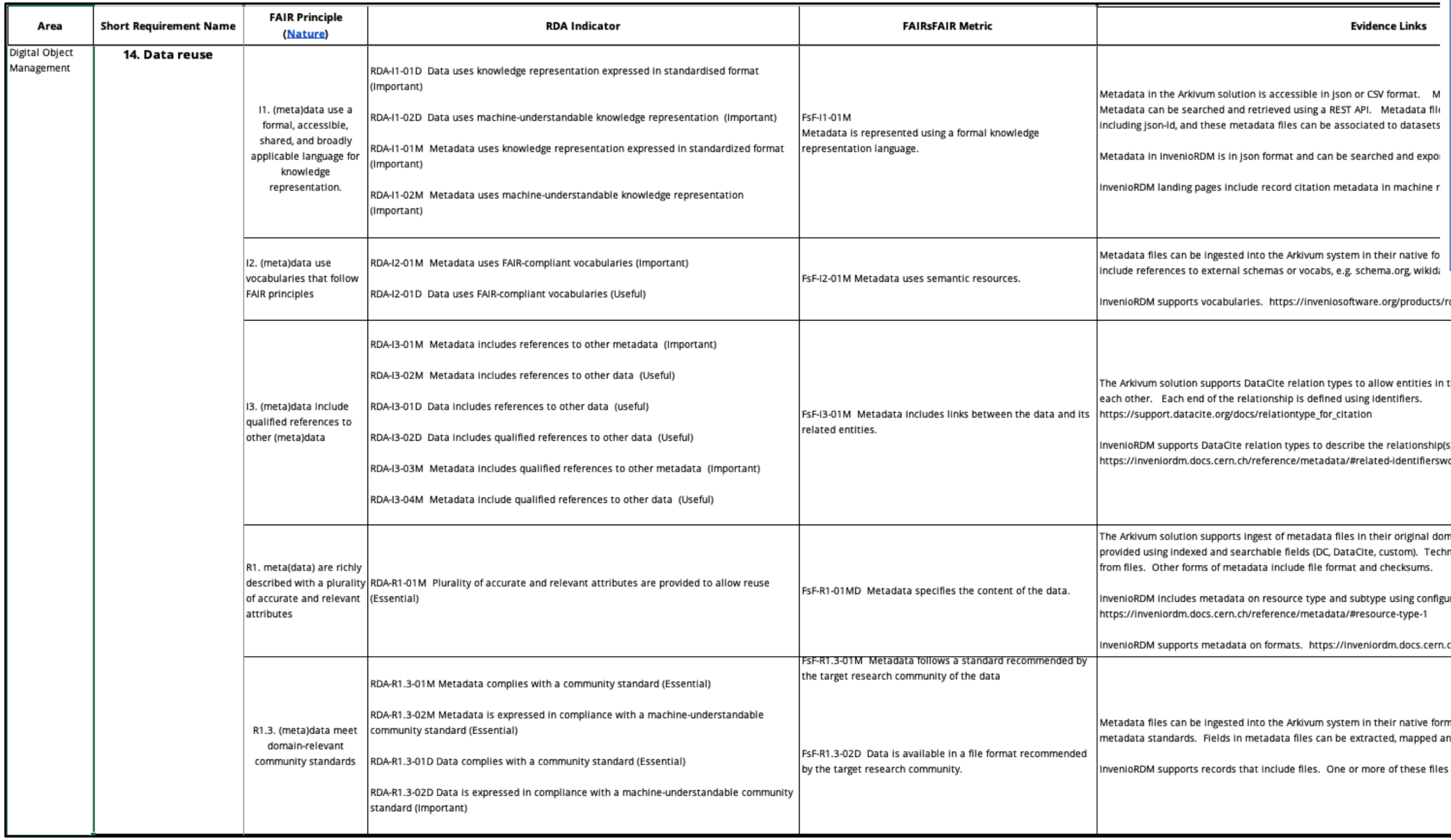
Publisher: CERN Open Data Portal

Rights

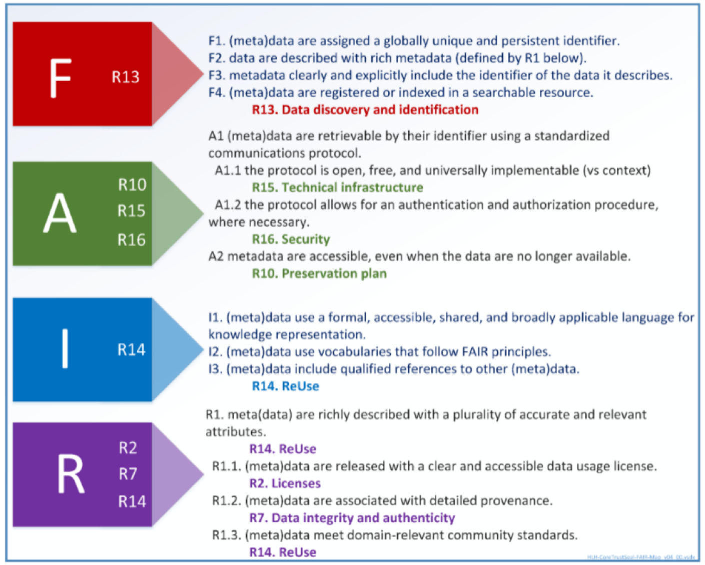
Creative Commons Zero v1.0 Universal

Export

CoreTrustSeal+FAIRenabling CapMat



Area	Short Requirement Name	FAIR Principle (Nature)	RDA Indicator	FAIRsFAIR Metric	Evidence Links
Digital Object Management	**14. Data reuse**	I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.	RDA-I1-01D Data uses knowledge representation expressed in standardised format (Important) RDA-I1-02D Data uses machine-understandable knowledge representation (Important) RDA-I1-01M Metadata uses knowledge representation expressed in standardized format (Important) RDA-I1-02M Metadata uses machine-understandable knowledge representation (Important)	FsF-I1-01M Metadata is represented using a formal knowledge representation language.	Metadata in the Arkivum solution is accessible in json or CSV format. M Metadata can be searched and retrieved using a REST API. Metadata fil including json-ld, and these metadata files can be associated to datasets Metadata in InvenioRDM is in json format and can be searched and expo InvenioRDM landing pages include record citation metadata in machine r
		I2. (meta)data use vocabularies that follow FAIR principles	RDA-I2-01M Metadata uses FAIR-compliant vocabularies (Important) RDA-I2-01D Data uses FAIR-compliant vocabularies (Useful)	FsF-I2-01M Metadata uses semantic resources.	Metadata files can be ingested into the Arkivum system in their native fo include references to external schemas or vocabs, e.g. schema.org, wikid InvenioRDM supports vocabularies. https://inveniosoftware.org/products/rdm/roadmap/
		I3. (meta)data include qualified references to other (meta)data	RDA-I3-01M Metadata includes references to other metadata (Important) RDA-I3-02M Metadata includes references to other data (Useful) RDA-I3-01D Data includes references to other data (useful) RDA-I3-02D Data includes qualified references to other data (Useful) RDA-I3-03M Metadata includes qualified references to other metadata (Important) RDA-I3-04M Metadata include qualified references to other data (Useful)	FsF-I3-01M Metadata includes links between the data and its related entities.	The Arkivum solution supports DataCite relation types to allow entities in the system (files, datasets etc.) to be related to each other. Each end of the relationship is defined using identifiers. https://support.datacite.org/docs/relationtype_for_citation InvenioRDM supports DataCite relation types to describe the relationship(s) of a record to other resources. https://inveniordm.docs.cern.ch/reference/metadata/#related-identifiersworks-0-n
		R1. meta(data) are richly described with a plurality of accurate and relevant attributes	RDA-R1-01M Plurality of accurate and relevant attributes are provided to allow reuse (Essential)	FsF-R1-01MD Metadata specifies the content of the data.	The Arkivum solution supports ingest of metadata files in their original domain specific format. Metadata can also be provided using indexed and searchable fields (DC, DataCite, custom). Technical metadata can be automatically extracted from files. Other forms of metadata include file format and checksums. InvenioRDM includes metadata on resource type and subtype using configurable controlled vocabs. https://inveniordm.docs.cern.ch/reference/metadata/#resource-type-1 InvenioRDM supports metadata on formats. https://inveniordm.docs.cern.ch/reference/metadata/#formats-0-n
		R1.3. (meta)data meet domain-relevant community standards	RDA-R1.3-01M Metadata complies with a community standard (Essential) RDA-R1.3-02M Metadata is expressed in compliance with a machine-understandable community standard (Essential) RDA-R1.3-01D Data complies with a community standard (Essential) RDA-R1.3-02D Data is expressed in compliance with a machine-understandable community standard (Important)	FsF-R1.3-01M Metadata follows a standard recommended by the target research community of the data FsF-R1.3-02D Data is available in a file format recommended by the target research community.	Metadata files can be ingested into the Arkivum system in their native format (XML or json), e.g. in domain specific metadata standards. Fields in metadata files can be extracted, mapped and indexed so metadata is searchable. InvenioRDM supports records that include files. One or more of these files could be a domain specific metadata format.



F1. (meta)data are assigned a globally unique and persistent identifier.
F2. data are described with rich metadata (defined by R1 below).
F3. metadata clearly and explicitly include the identifier of the data it describes.
F4. (meta)data are registered or indexed in a searchable resource.
R13. Data discovery and identification

A1 (meta)data are retrievable by their identifier using a standardized communications protocol.
A1.1 the protocol is open, free, and universally implementable (vs context)
R15. Technical infrastructure
A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
R16. Security
A2 metadata are accessible, even when the data are no longer available.
R10. Preservation plan

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles.
I3. (meta)data include qualified references to other (meta)data.
R14. ReUse

R1. meta(data) are richly described with a plurality of accurate and relevant attributes.
R14. ReUse
R1.1. (meta)data are released with a clear and accessible data usage license.
R2. Licenses
R1.2. (meta)data are associated with detailed provenance.
R7. Data integrity and authenticity
R1.3. (meta)data meet domain-relevant community standards.
R14. ReUse

CORE TRUST SEAL

FAIRSFAIR
Fostering Fair Data Practices in Europe

Cloud-hosted Processing of Archived Research Datasets

- GCP hosted applications
  - Cloud Functions (serverless code)
  - Cloud Run (containers)
  - GKE (Kubernetes)
  - Compute Engine (VMs)
- Arkivum Webhooks
  - Ingest, preservation, access
- Arkivum REST API
  - including search, get metadata, export files
- Xrootd server
  - Easy integration of scientific apps

Portability

- Single software stack that can run on AWS, GCP and OpenStack
- Deployed on GCP and at CERN for ARCHIVER
- Ingest and export of data and metadata in native formats
- Interact with solution using standard protocols (s3 buckets, REST API, eduGAIN, Web UI)

Google Cloud

aws

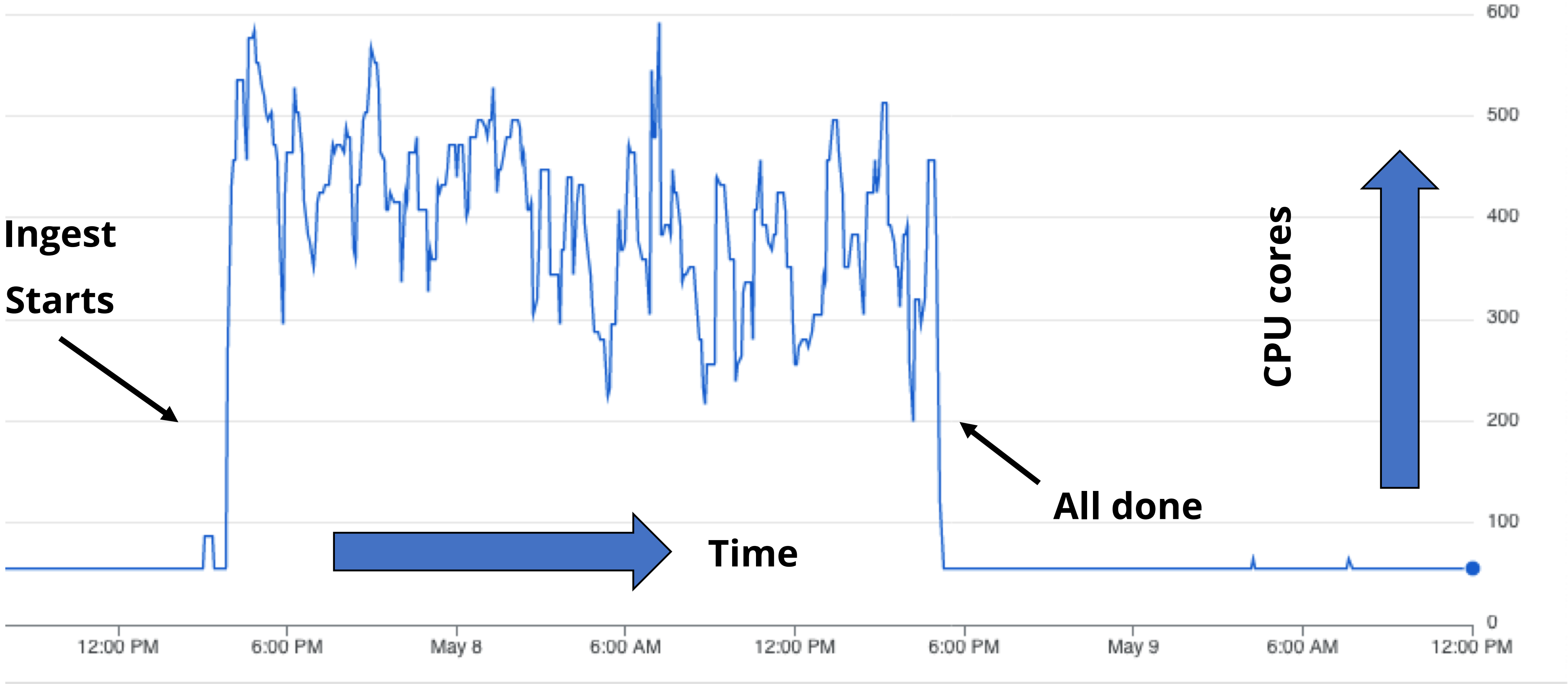
openstack

Scalability and Performance

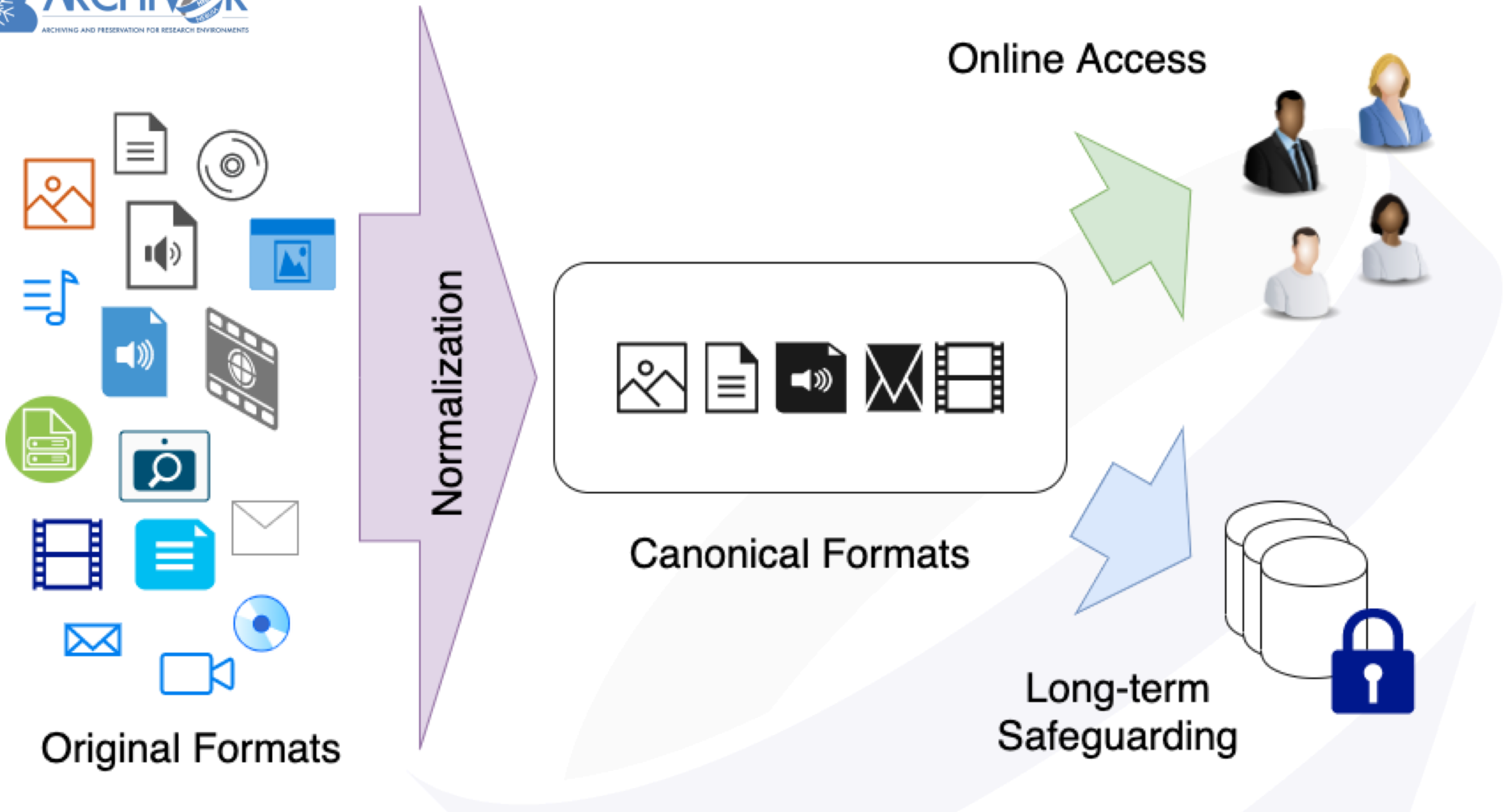
Serverless Computing: Scalability, Performance, Efficiency

- Scalable and cost-effective archiving workflows and processing
  - Kubernetes and autoscaling
  - Scale to zero as well as autoscaling for peak loads (pods and nodes)
  - Pre-emptible nodes to reduce costs (up to 70% lower)
  - Terraform and Ansible for provisioning
  - Rancher, Prometheus, Grafana, Kibana for monitoring and analytics
- Microservices
  - Checksum, virus scan, file format identification, caching, replication, unpack …
  - Stateless and able to run in parallel (can 100TB+ per day)
  - Jobs recorded and tracked

Example: Ingest and Archiving of Astronomy Datasets



440,000 image files (25TB) ingested as 2780 big data bags within 24hrs



Online Access

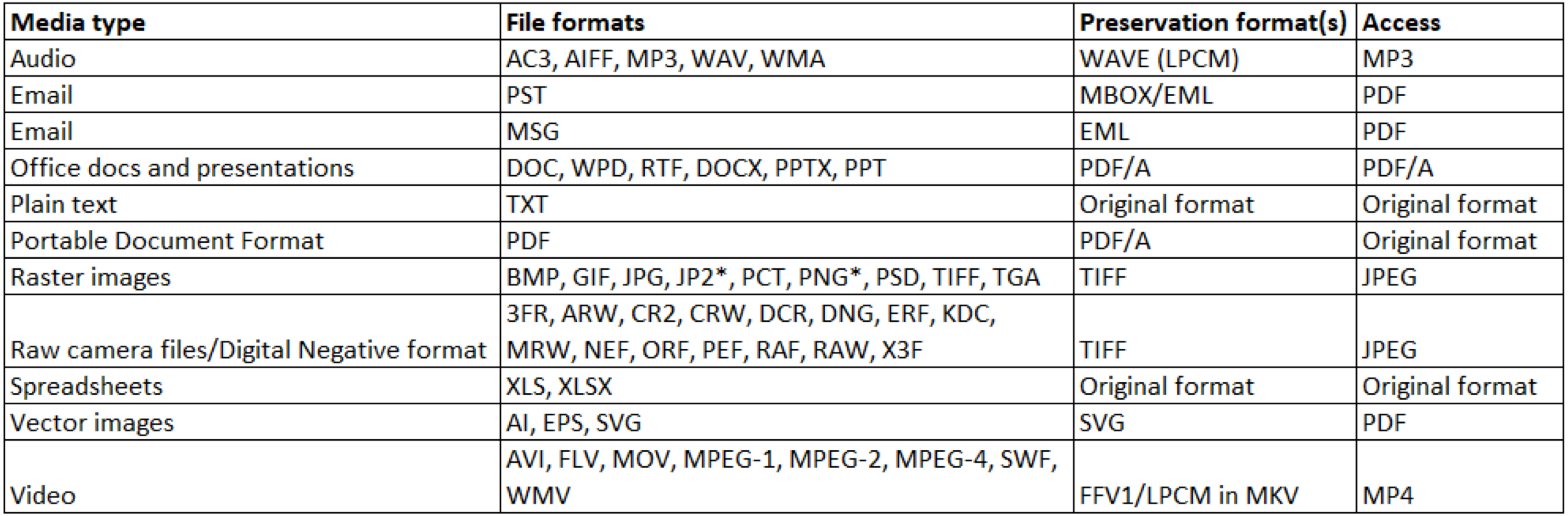
Normalization

Canonical Formats

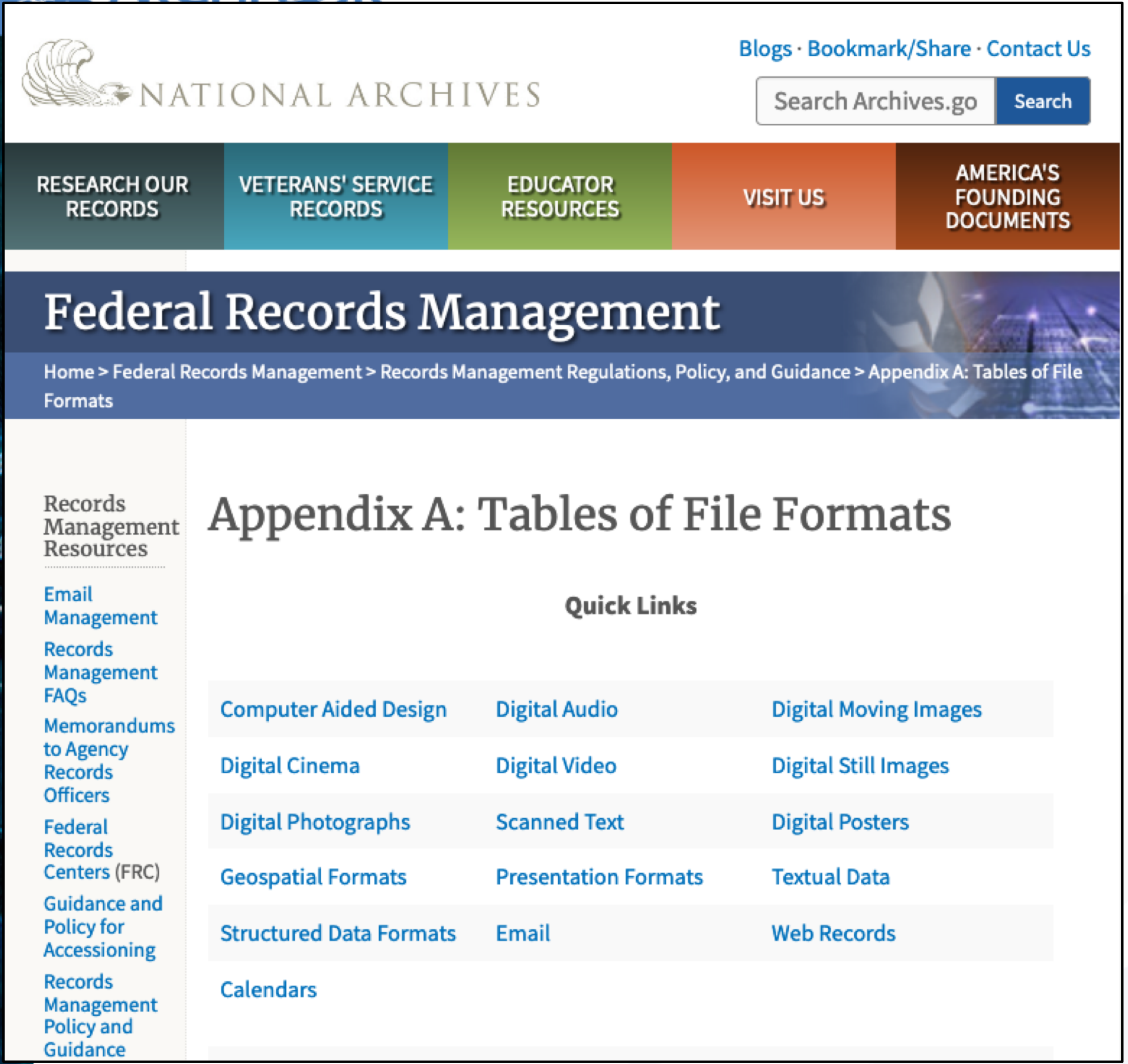
Long-term Safeguarding

Original Formats

Preservation Formats and Access Formats



Media type	File formats	Preservation format(s)	Access
Audio	AC3, AIFF, MP3, WAV, WMA	WAVE (LPCM)	MP3
Email	PST	MBOX/EML	PDF
Email	MSG	EML	PDF
Office docs and presentations	DOC, WPD, RTF, DOCX, PPTX, PPT	PDF/A	PDF/A
Plain text	TXT	Original format	Original format
Portable Document Format	PDF	PDF/A	Original format
Raster images	BMP, GIF, JPG, JP2*, PCT, PNG*, PSD, TIFF, TGA	TIFF	JPEG
Raw camera files/Digital Negative format	3FR, ARW, CR2, CRW, DCR, DNG, ERF, KDC, MRW, NEF, ORF, PEF, RAF, RAW, X3F	TIFF	JPEG
Spreadsheets	XLS, XLSX	Original format	Original format
Vector images	AI, EPS, SVG	SVG	PDF
Video	AVI, FLV, MOV, MPEG-1, MPEG-2, MPEG-4, SWF, WMV	FFV1/LPCM in MKV	MP4


NATIONAL ARCHIVES
Blogs · Bookmark/Share · Contact Us
Search Archives.go
Search
RESEARCH OUR RECORDS
VETERANS' SERVICE RECORDS
EDUCATOR RESOURCES
VISIT US
AMERICA'S FOUNDING DOCUMENTS
Records Management Resources
Email Management
Records Management FAQs
Memorandums to Agency Records Officers
Federal Records Centers (FRC)
Guidance and Policy for Accessioning
Records Management Policy and Guidance

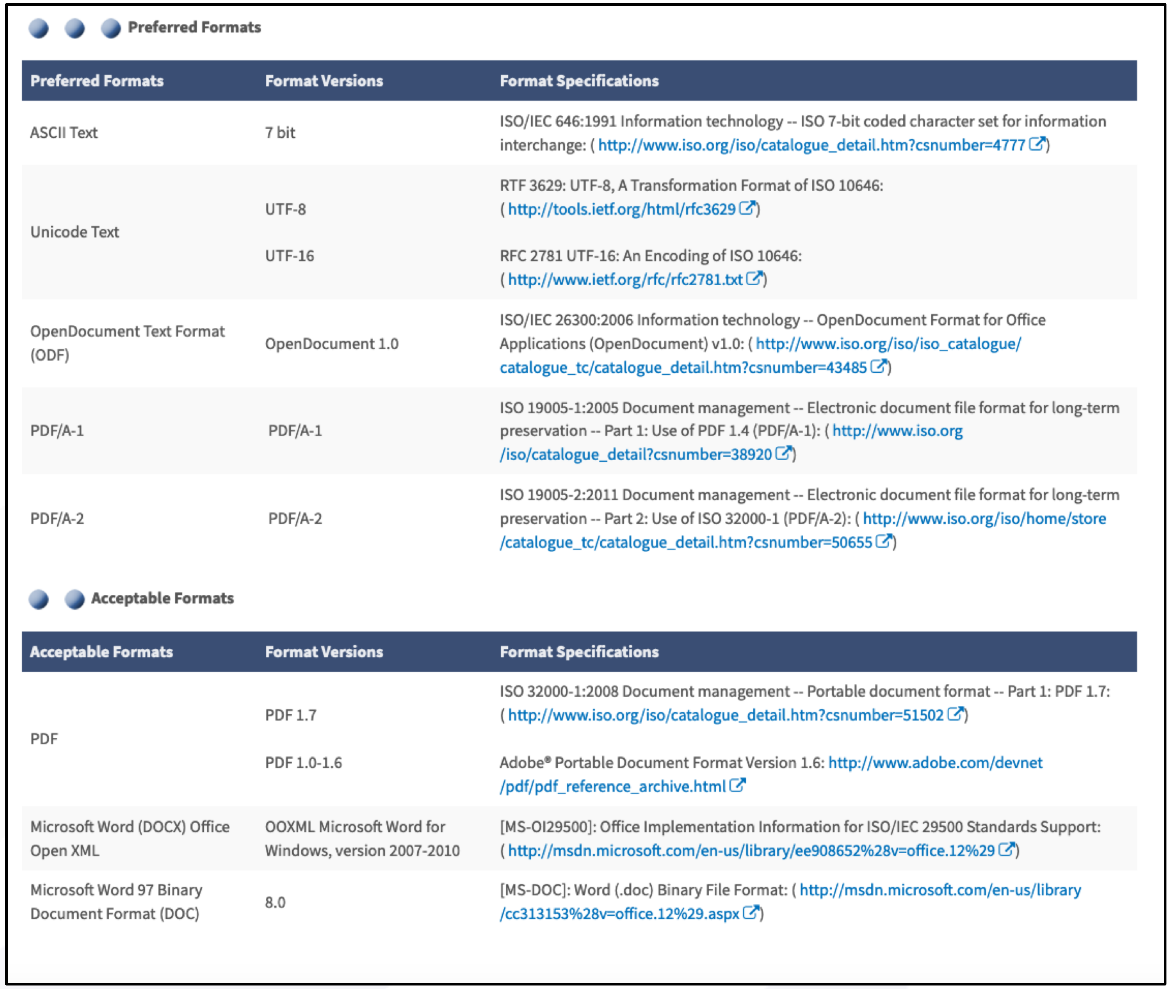
Federal Records Management

Home > Federal Records Management > Records Management Regulations, Policy, and Guidance > Appendix A: Tables of File Formats

Appendix A: Tables of File Formats

Quick Links

Computer Aided Design	Digital Audio	Digital Moving Images
Digital Cinema	Digital Video	Digital Still Images
Digital Photographs	Scanned Text	Digital Posters
Geospatial Formats	Presentation Formats	Textual Data
Structured Data Formats	Email	Web Records
Calendars		



●●● **Preferred Formats**

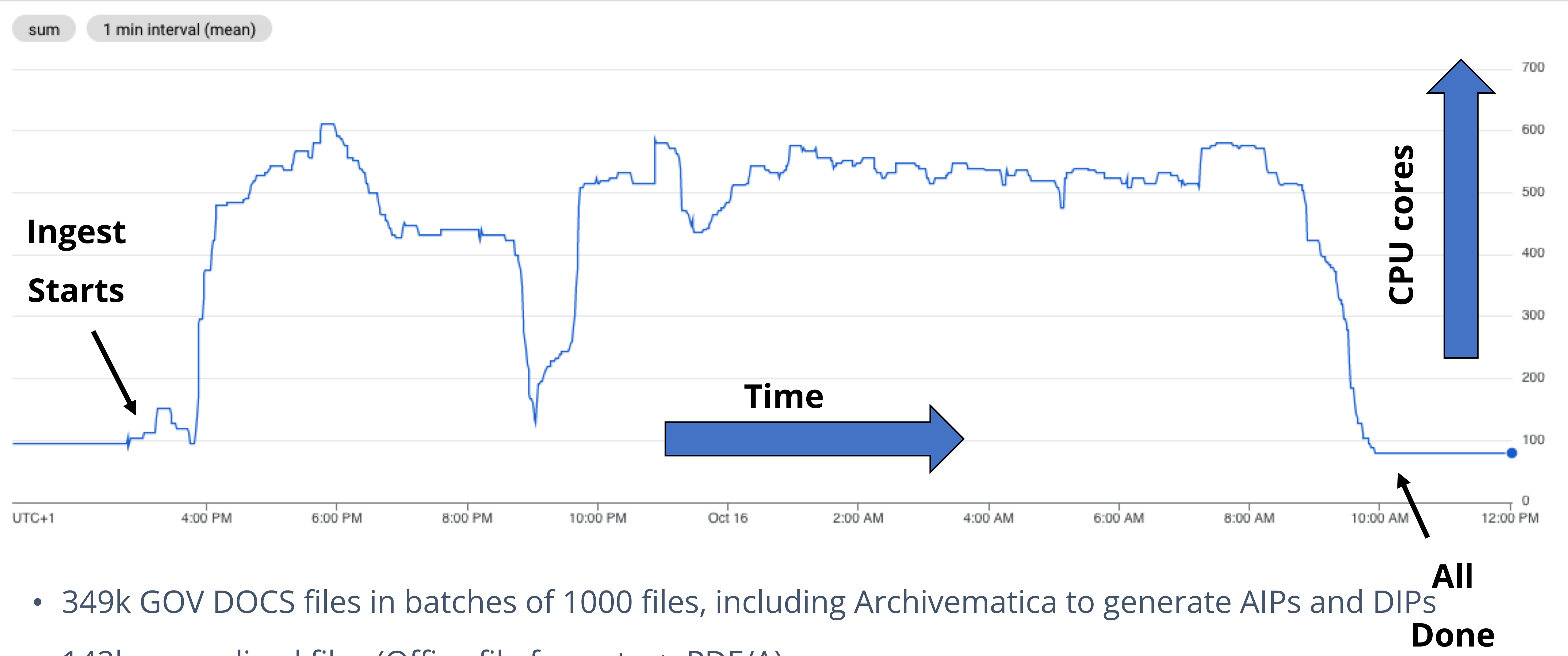
Preferred Formats	Format Versions	Format Specifications
ASCII Text	7 bit	ISO/IEC 646:1991 Information technology -- ISO 7-bit coded character set for information interchange: (http://www.iso.org/iso/catalogue_detail.htm?csnumber=4777)
Unicode Text	UTF-8	RTF 3629: UTF-8, A Transformation Format of ISO 10646: (http://tools.ietf.org/html/rfc3629)
	UTF-16	RFC 2781 UTF-16: An Encoding of ISO 10646: (http://www.ietf.org/rfc/rfc2781.txt)
OpenDocument Text Format (ODF)	OpenDocument 1.0	ISO/IEC 26300:2006 Information technology -- OpenDocument Format for Office Applications (OpenDocument) v1.0: (http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=43485)
PDF/A-1	PDF/A-1	ISO 19005-1:2005 Document management -- Electronic document file format for long-term preservation -- Part 1: Use of PDF 1.4 (PDF/A-1): (http://www.iso.org/iso/catalogue_detail?csnumber=38920)
PDF/A-2	PDF/A-2	ISO 19005-2:2011 Document management -- Electronic document file format for long-term preservation -- Part 2: Use of ISO 32000-1 (PDF/A-2): (http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=50655)

●● **Acceptable Formats**

Acceptable Formats	Format Versions	Format Specifications
PDF	PDF 1.7	ISO 32000-1:2008 Document management -- Portable document format -- Part 1: PDF 1.7: (http://www.iso.org/iso/catalogue_detail.htm?csnumber=51502)
	PDF 1.0-1.6	Adobe® Portable Document Format Version 1.6: http://www.adobe.com/devnet/pdf/pdf_reference_archive.html
Microsoft Word (DOCX) Office Open XML	OOXML Microsoft Word for Windows, version 2007-2010	[MS-OI29500]: Office Implementation Information for ISO/IEC 29500 Standards Support: (http://msdn.microsoft.com/en-us/library/ee908652%28v=office.12%29)
Microsoft Word 97 Binary Document Format (DOC)	8.0	[MS-DOC]: Word (.doc) Binary File Format: (http://msdn.microsoft.com/en-us/library/cc313153%28v=office.12%29.aspx)

https://www.archives.gov/records-mgmt/policy/transfer-guidance-tables.html

Example: Preserving Documents



- 349k GOV DOCS files in batches of 1000 files, including Archivematica to generate AIPs and DIPs
- 142k normalised files (Office file formats -> PDF/A)

Leveraging the Scalability and Efficiency of Cloud Infrastructure

POLICY FORUM

As demand for data centers rises, energy efficiency improvements to the IT devices and cooling systems they house can keep energy use in check.

ENERGY

Recalibrating global data center energy-use estimates

Growth in energy use has slowed owing to efficiency gains that smart policies can help maintain in the near term

By **Eric Masanet[1,2], Arman Shehabi[3], Nuoa Lei[1], Sarah Smith[3], Jonathan Koomey[4]**

Data centers represent the information backbone of an increasingly digitalized world. Demand for their services has been rising rapidly (*1*), and data-intensive technologies such as artificial intelligence, smart and connected energy systems, distributed manufacturing systems, and autonomous vehicles promise to increase demand further (*2*). Given that data centers are energy-intensive enterprises, estimated to account for around 1% of worldwide electricity use, these trends have clear implications for global energy demand and must be analyzed rigorously. Several oft-cited yet simplistic analyses claim that the energy used by the world's data centers has doubled over the past decade and that their energy demand for data center services rises rapidly, so too must their global energy use. But such extrapolations based on recent service demand growth indicators overlook strong countervailing energy efficiency trends that have occurred in parallel (see the first figure). Here, we integrate new data from different sources that have emerged recently and suggest more modest growth in global data center energy use (see the second figure). This provides policy-makers and energy analysts a recalibrated understanding of global data center energy use, its drivers, and near-term efficiency potential.

Assessing implications of growing demand for data centers requires robust understanding of the scale and drivers of global data center energy use that has eluded many policy-makers and energy analysts. The reason for this blind spot is a historical lack of "bottom-up" information Bottom-up analyses tend to best reflect this broad range of factors, generating the most credible historical and near-term energy-use estimates (*7*). Despite several recent national studies (*8*), the latest fully replicable bottom-up estimates of global data center energy use appeared nearly a decade ago. These estimates suggested that the worldwide energy use of data centers had grown from 153 terawatt-hours (TWh) in 2005 to between 203 and 273 TWh by 2010, totaling 1.1 to 1.5% of global electricity use (*9*).

Since 2010, however, the data center landscape has changed dramatically (see the first figure). By 2018, global data center workloads and compute instances had increased more than sixfold, whereas data center internet protocol (IP) traffic had increased by more than 10-fold (*1*). Data center storage capacity has also grown rapidly, increasing by an estimated factor of 25 over the same time period (*1*, *8*). There has been a tendency among analysts to use such service demand trends to simply extrapolate earlier bottom-up energy values, leading to unreliable predictions of current and future global data center energy use (*3*–*5*). They might, for example, scale up previous bottom-up values (e.g., total data center energy use in 2010) on the basis of the growth rate of a service demand indicator (e.g., growth in global IP traffic from 2010 to 2020) to arrive at an estimate of future energy use (e.g., total data center energy use in 2020).

But since 2010, electricity use per computation of a typical volume server—the workhorse of the data center—has dropped by a factor of four, largely owing to processor-efficiency improvements and reductions in idle power (*10*). At the same time, the watts per terabyte of installed storage has dropped by an estimated factor of nine owing to storage-drive density and efficiency gains (*8*). Furthermore, growth in the number of servers has slowed considerably owing to a fivefold increase in the average number of compute instances hosted per server (owing to virtualization), alongside steady reductions in data center power usage effectiveness (PUE, the total amount

https://datacenters.lbl.gov/sites/default/files/Masanet_et_al_Science_2020.full_.pdf

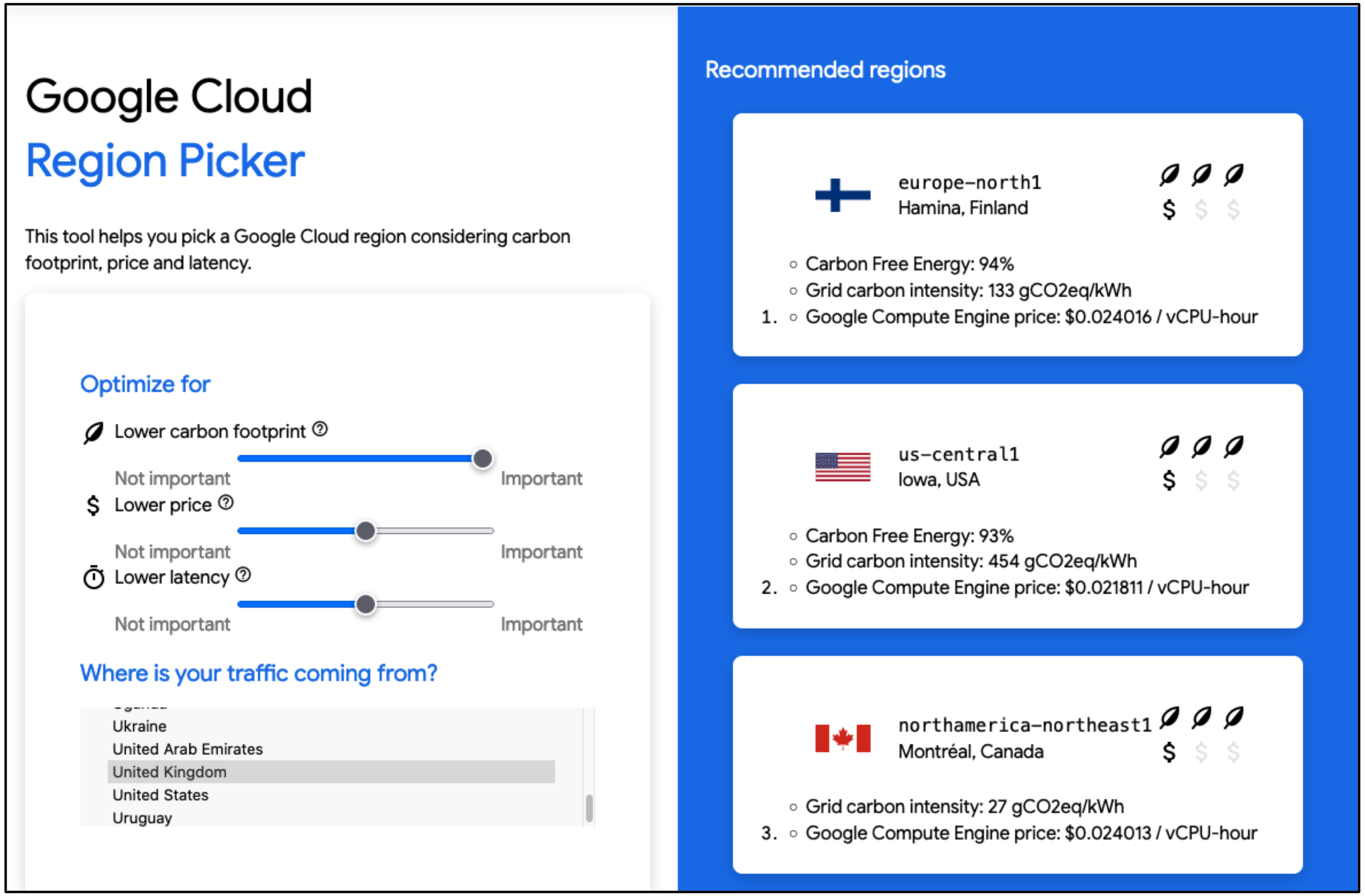
KEDA

RANCHER

ANSIBLE

Terraform

Green Data Centers, Renewable Energy



Google Cloud

Region Picker

This tool helps you pick a Google Cloud region considering carbon footprint, price and latency.

Optimize for

Lower carbon footprint

Not important Important

$ Lower price

Not important Important

Lower latency

Not important Important

Where is your traffic coming from?

Uganda
Ukraine
United Arab Emirates
United Kingdom
United States
Uruguay

Recommended regions

1. europe-north1
   Hamina, Finland
   $
   - Carbon Free Energy: 94%
   - Grid carbon intensity: 133 gCO2eq/kWh
   - Google Compute Engine price: $0.024016 / vCPU-hour

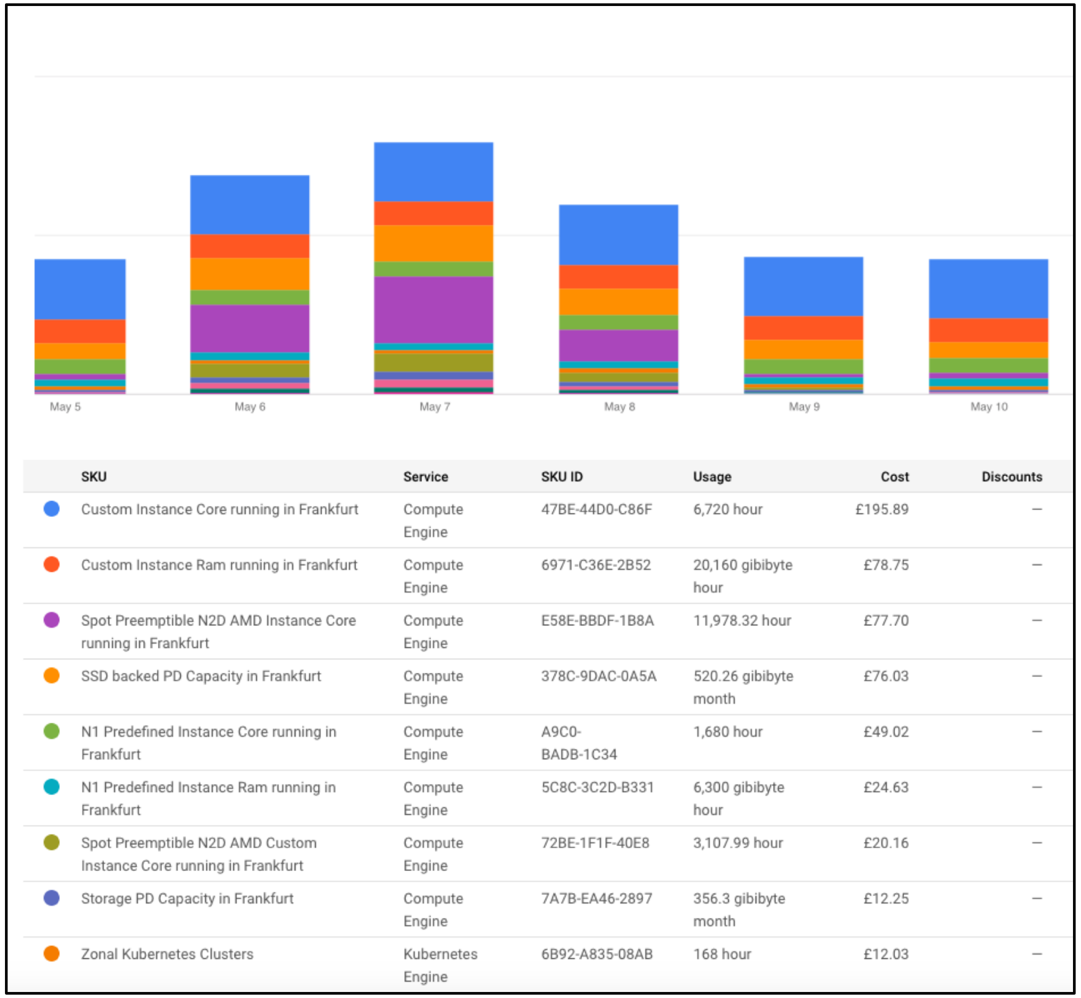
2. us-central1
   Iowa, USA
   $
   - Carbon Free Energy: 93%
   - Grid carbon intensity: 454 gCO2eq/kWh
   - Google Compute Engine price: $0.021811 / vCPU-hour

3. northamerica-northeast1
   Montréal, Canada
   $
   - Grid carbon intensity: 27 gCO2eq/kWh
   - Google Compute Engine price: $0.024013 / vCPU-hour

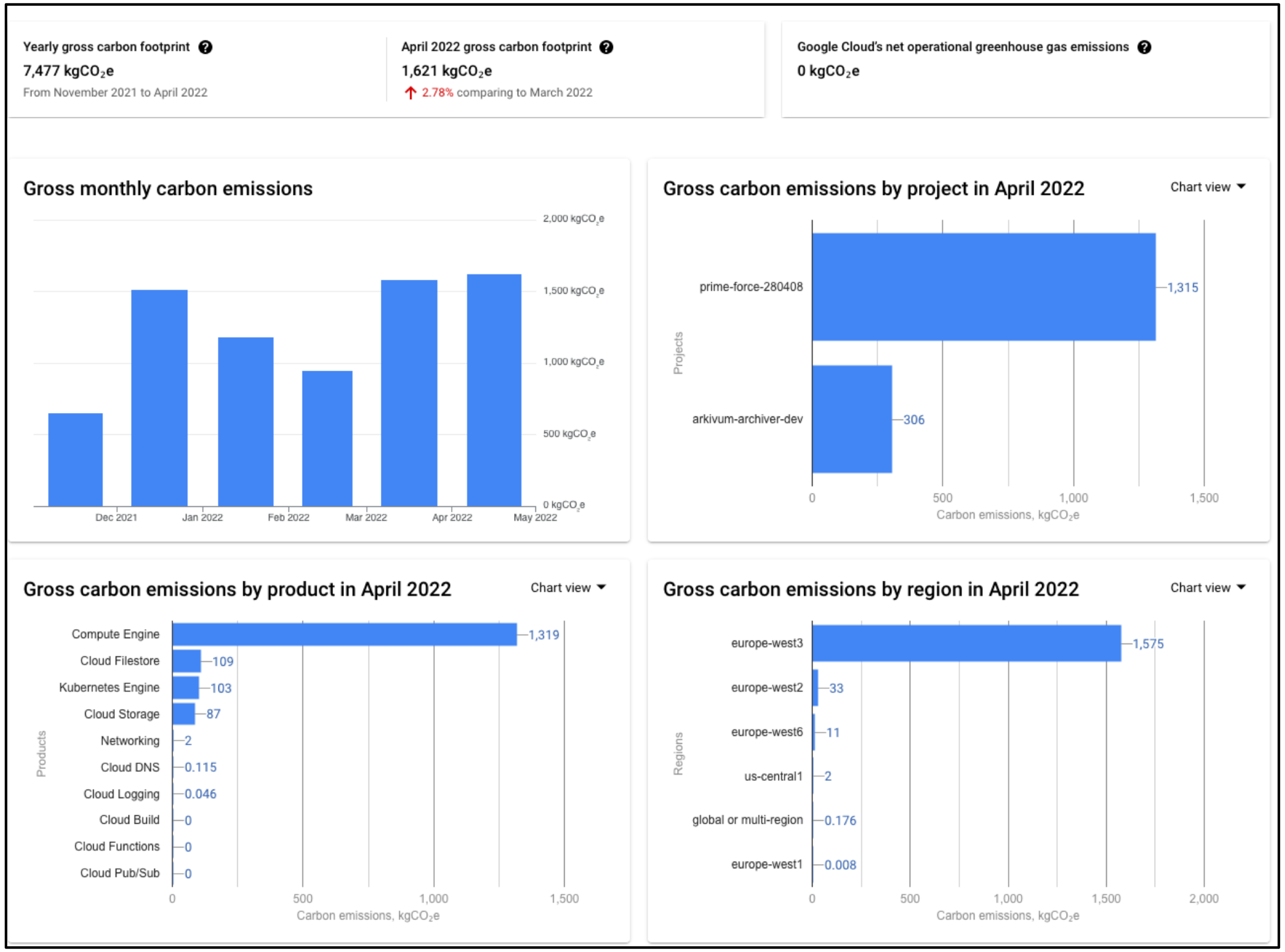
https://cloud.withgoogle.com/region-picker/

https://www.google.com/about/datacenters/gallery/#hamina-exterior-landscape

Costs, Resource Consumption, Carbon Footprint



	SKU	Service	SKU ID	Usage	Cost	Discounts
●	Custom Instance Core running in Frankfurt	Compute Engine	47BE-44D0-C86F	6,720 hour	£195.89	–
●	Custom Instance Ram running in Frankfurt	Compute Engine	6971-C36E-2B52	20,160 gibibyte hour	£78.75	–
●	Spot Preemptible N2D AMD Instance Core running in Frankfurt	Compute Engine	E58E-BBDF-1B8A	11,978.32 hour	£77.70	–
●	SSD backed PD Capacity in Frankfurt	Compute Engine	378C-9DAC-0A5A	520.26 gibibyte month	£76.03	–
●	N1 Predefined Instance Core running in Frankfurt	Compute Engine	A9C0-BADB-1C34	1,680 hour	£49.02	–
●	N1 Predefined Instance Ram running in Frankfurt	Compute Engine	5C8C-3C2D-B331	6,300 gibibyte hour	£24.63	–
●	Spot Preemptible N2D AMD Custom Instance Core running in Frankfurt	Compute Engine	72BE-1F1F-40E8	3,107.99 hour	£20.16	–
●	Storage PD Capacity in Frankfurt	Compute Engine	7A7B-EA46-2897	356.3 gibibyte month	£12.25	–
●	Zonal Kubernetes Clusters	Kubernetes Engine	6B92-A835-08AB	168 hour	£12.03	–



Yearly gross carbon footprint ?
7,477 kgCO$_2$e
From November 2021 to April 2022

April 2022 gross carbon footprint ?
1,621 kgCO$_2$e
↑ 2.78% comparing to March 2022

Google Cloud's net operational greenhouse gas emissions ?
0 kgCO$_2$e

Gross monthly carbon emissions

Gross carbon emissions by project in April 2022

Gross carbon emissions by product in April 2022

Gross carbon emissions by region in April 2022

Benchmarking and Metrics

- Execute real world scenarios
- Record parameters
  - execution time
  - data volumes, number of files
  - type of activity (ingest, export, preservation)
- Extract costs and resource consumption from cloud provider
- Extract carbon footprint from cloud provider
- Add short-term and long-term storage
  - Upload/export buckets, caching, archive buckets
- Calculate metrics

Ingest	**$2.2 per TB**
	0.1 kgCO$_2$eq
Long-term storage	**$30 per TB-year**
	0.7 kgCO$_2$eq per TB-year

Summary

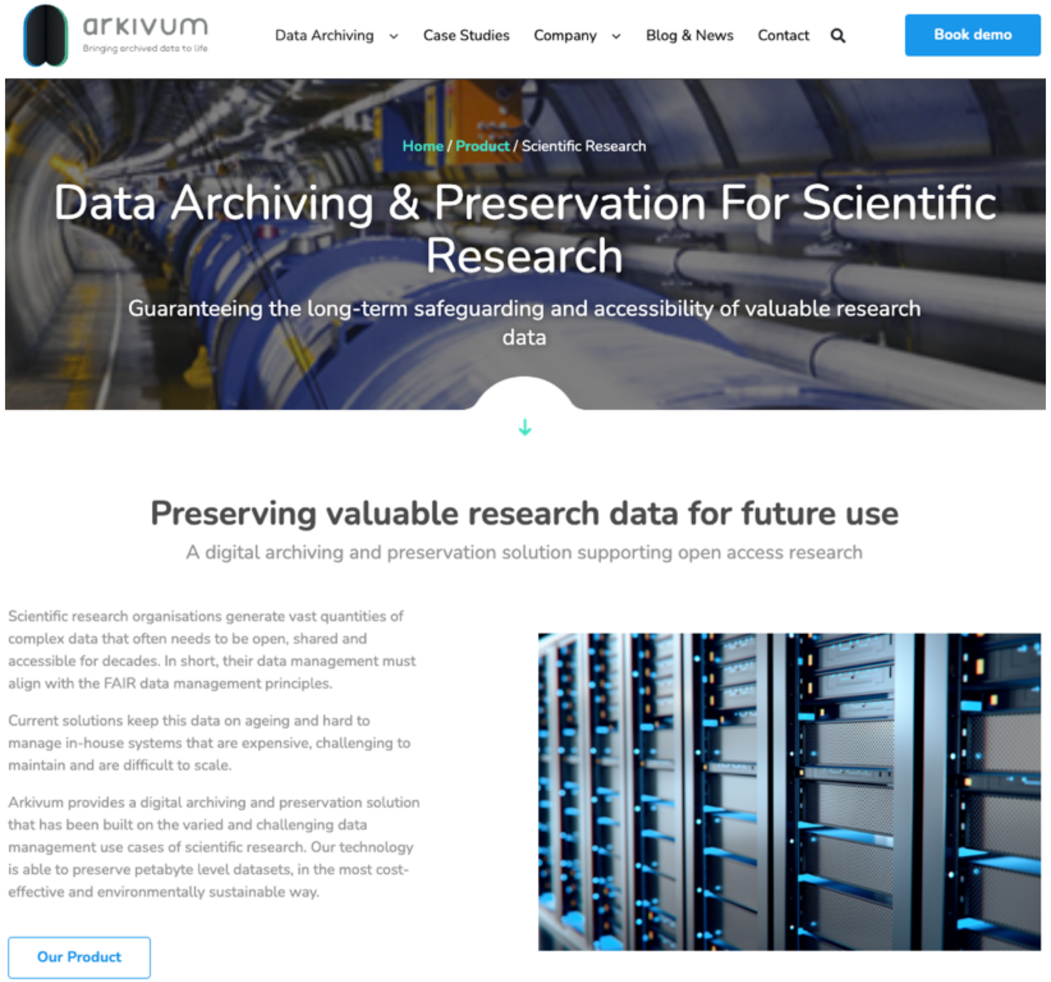
The Arkivum Solution for ARCHIVER

- Highly scalable LTDP capable of ingesting and preserving 100TB+ per day
- Co-locate scientific applications with archived data
- Integration with InvenioRDM for creating/publishing landing pages
- Serverless computing: only consume what's needed and when it's needed
- Cost-efficient and minimized carbon-footprint
- Deployment using GCP, AWS and on-premise
- Provided as a fully managed service / SaaS solution
- Supports LTDP requirements and models (DPC RAM and NDSA levels)
- Supports TDR and FAIR requirements and models (CoreTrustSeal and FAIR)

Arkivum Commercialisation Plans

Arkivum's Commercial Credentials

- Operating since 2011 - transitioned from predominantly hardware to software business
- Customer base spanning Higher Education, Heritage, Corporate and Life Sciences organisations
- 50+ customers
- One platform/product
  - ARCHIVER work has been introduced into the core product outside of the scope of the three Phases.
  - Annual recurring cost – options available.
- Commercialisation plan delivered as part of Pilot Phase.



Data Archiving & Preservation For Scientific Research

Guaranteeing the long-term safeguarding and accessibility of valuable research data

Preserving valuable research data for future use

A digital archiving and preservation solution supporting open access research

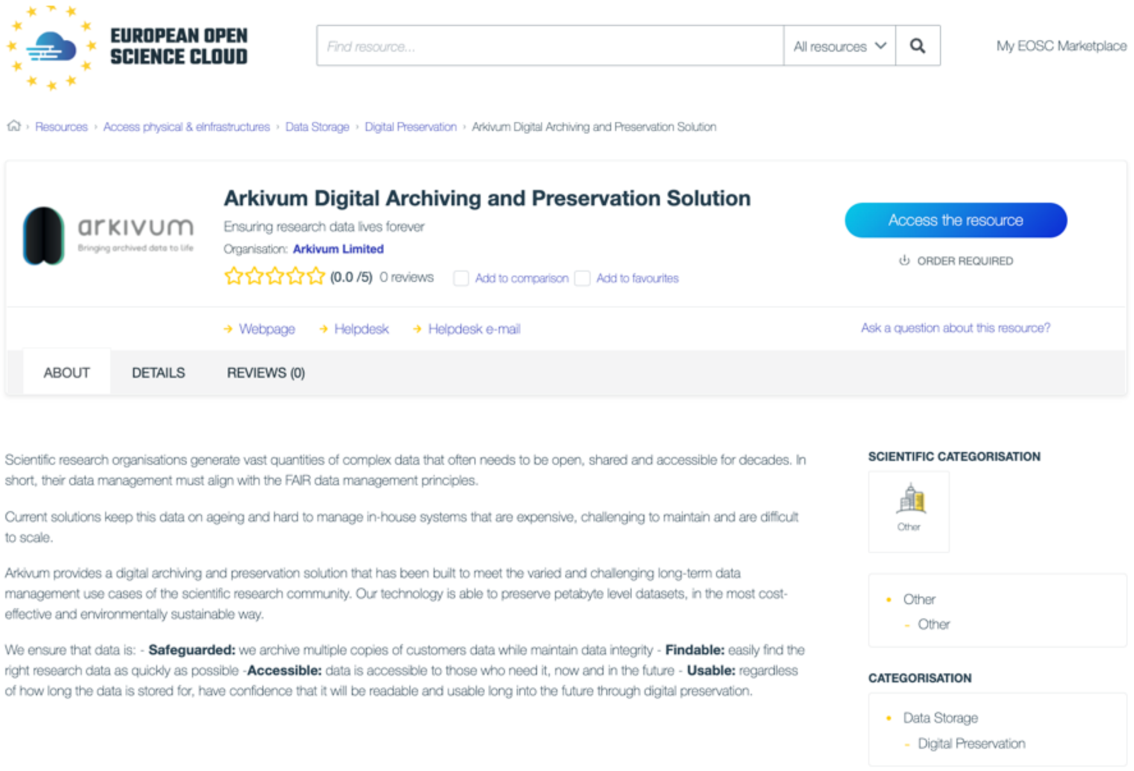
Scientific research organisations generate vast quantities of complex data that often needs to be open, shared and accessible for decades. In short, their data management must align with the FAIR data management principles.

Current solutions keep this data on ageing and hard to manage in-house systems that are expensive, challenging to maintain and are difficult to scale.

Arkivum provides a digital archiving and preservation solution that has been built on the varied and challenging data management use cases of scientific research. Our technology is able to preserve petabyte level datasets, in the most cost-effective and environmentally sustainable way.

1. Future European Commission Projects and EOSC

- EOSC Marketplace listing
  - Early Adopters Programme
- Ordering possible through EOSC (via the Arkivum website)
- Active involvement in past, present and future events
- Support future initiatives and projects


EUROPEAN OPEN SCIENCE CLOUD
Find resource...
All resources
My EOSC Marketplace
Resources › Access physical & eInfrastructures › Data Storage › Digital Preservation › Arkivum Digital Archiving and Preservation Solution
arkivum
Bringing archived data to life
Access the resource
ORDER REQUIRED
(0.0 /5) 0 reviews
□ Add to comparison
□ Add to favourites
Webpage
Helpdesk
Helpdesk e-mail
Ask a question about this resource?
ABOUT
DETAILS
REVIEWS (0)
Other

Arkivum Digital Archiving and Preservation Solution

Ensuring research data lives forever

Organisation: **Arkivum Limited**

Scientific research organisations generate vast quantities of complex data that often needs to be open, shared and accessible for decades. In short, their data management must align with the FAIR data management principles.

Current solutions keep this data on ageing and hard to manage in-house systems that are expensive, challenging to maintain and are difficult to scale.

Arkivum provides a digital archiving and preservation solution that has been built to meet the varied and challenging long-term data management use cases of the scientific research community. Our technology is able to preserve petabyte level datasets, in the most cost-effective and environmentally sustainable way.

We ensure that data is: - **Safeguarded:** we archive multiple copies of customers data while maintain data integrity - **Findable:** easily find the right research data as quickly as possible -**Accessible:** data is accessible to those who need it, now and in the future - **Usable:** regardless of how long the data is stored for, have confidence that it will be readable and usable long into the future through digital preservation.

SCIENTIFIC CATEGORISATION

- Other
  - Other

CATEGORISATION

- Data Storage
  - Digital Preservation

2. Direct Sales and Marketing

Content & webinars

Direct campaigns

Website and SEO

PR & Social

Events & conferences

3. Channel Partnerships

- Leverage existing (e.g. GCP) and new channel partnerships
- Expand reach
- Sales enablement and training
- Mutually beneficial commercial relationships

Google Cloud

Summary

- Arkivum's digital preservation and archiving solution is ready for market deployment today
- Proven track record of delivering commercial success within new markets
- Three main routes to market are:
    - Work with EC and EOSC in future projects
    - Direct sales and marketing
    - Build channel partnerships

arkivum
Bringing archived data to life

THANK YOU

Matthew Addis

Tom Lynam

www.arkivum.com | hello@arkivum.com

ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

Closing remarks

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.

Public Upcoming Events

- Training Webinars for Arkivum and Libnova
  - **Libnova**: Wednesday 22nd of June 2022; **Arkivum**: Thursday 23rd of June 2022
  - Information & Registration at: https://archiver-project.eu/all-events

- iPRES 2022 12th-16th to September: https://ipres2022.scot/
  - ARCHIVER Panel “*Sustainable Preservation of Scientific Data*”, on theme *Innovation*
  - Announcement of the winners of the DPC Awards 2022

- PV 2023: https://indico.cern.ch/event/838830/
  - To be hold in May 2023; Schedule to be announced

Acknowledgements

- Buyer Organisations
  - Antonio Vivace, Bob Jones, Ignacio Lozada, Jacek Chodak, Jakub Urban, Jamie Shiers, Jean-Yves Le Meur, Jordi Casals, Justin Clark-Casey, Manuel Delfino, Marc Riera, Mari-Carmen Porto, Martin Gasthuber, Sergey Yakubov, Steven Newhouse, Tibor Simko, Tony Wildish, Vaggelis Motesnitsalis, Volker Guelzow
- Project Office & Project Partners
  - Anna Manou, Anna Benzo, David Foster, Dominique Buyse, Florence Pesce, Jimmy James, Joelma Tolomeo, Josh Davison, Lucie Pocha, Mariletizia Mari, Marion Devouassoux, Miguel Santos, Ruben Van Caelenberg, Sara Pittonet Gaiarin, Sonia Mentrida, Viktor Varga
- External Advisory Board Members
  - Eberhard Mikusch, Ingrid Dillo, Jenny Mitcham, Nigel Houghton, Rudolf Dimper, Sarah Jones
- Bidders and Contractors from all Phases
- Early Adopter Organisations
- The European Commission & Panel of Reviewers

ARCHIVER

ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

Join our community!

www.archiver-project.eu

@ArchiverProject

company/archiver-project

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.