arkivum

Bringing archived data to life

The Arkivum Solution for ARCHIVER

Matthew Addis

Arkivum

Agenda

- Introduction (10 mins)
- Solution Overview (20 mins)
- Solution Training: Part 1 (30 mins)
- Break (5 mins)
- Solution Training: Part 2 (30 mins)
- Solution Training: Part 3 (40 mins)
- Questions and Discussion (30 mins)

Introduction (10 mins)

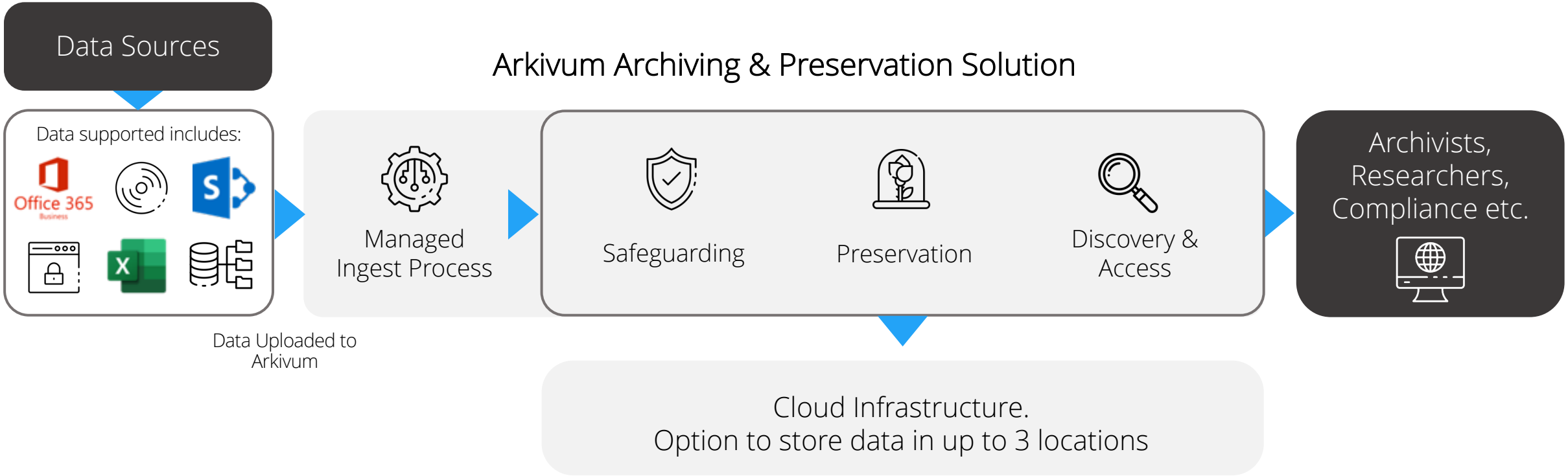
Agenda

- Arkivum Overview
- Product and Services
- Certification, Standards and Good Practice
- Approach to ARCHIVER
- Partnership with Google
- Availability of the solution in EOSC

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



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

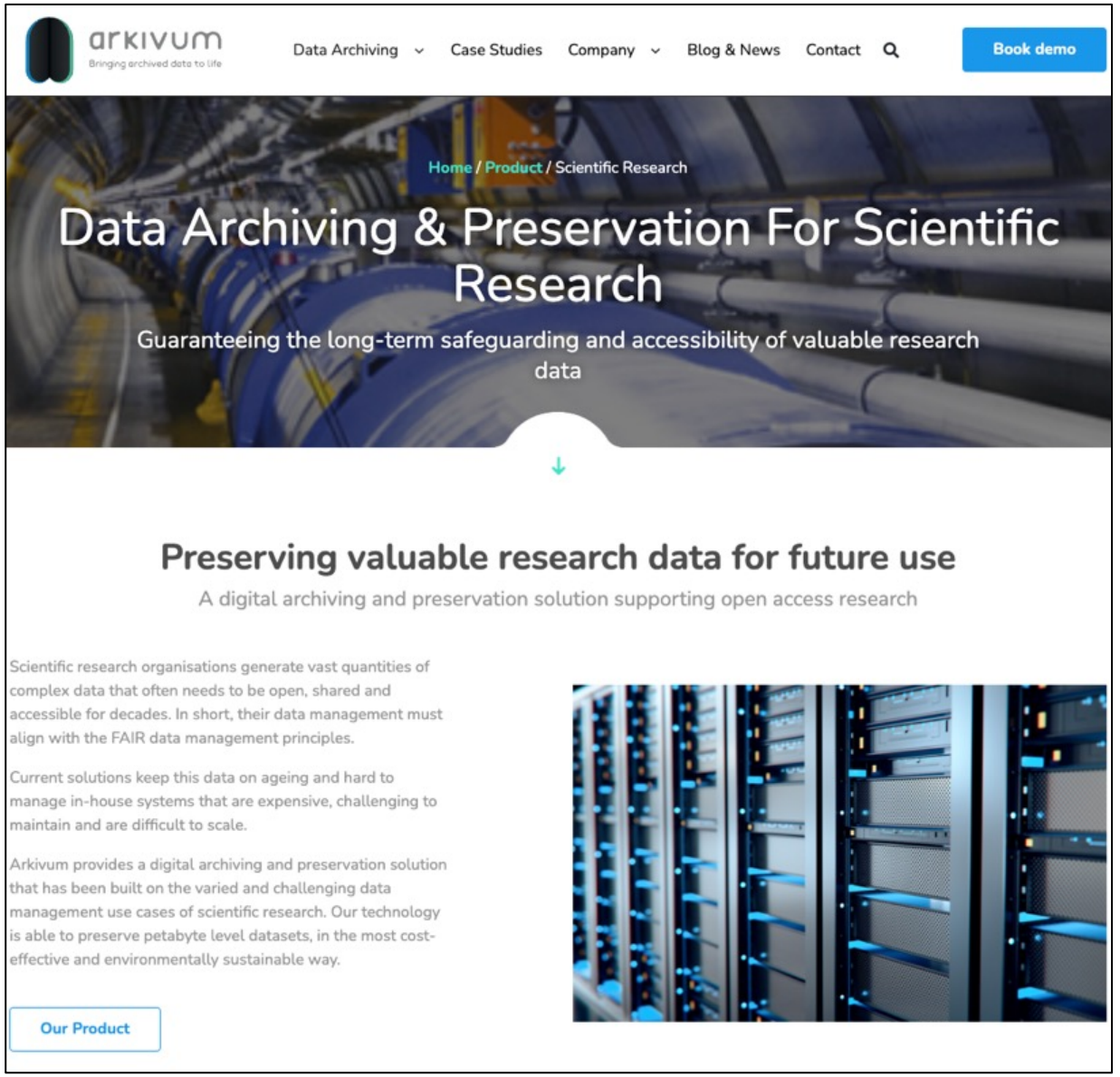
Services and Good Practice

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



Approach to ARCHIVER

- Collaborative and iterative development and testing
- Knowledge sharing across domains (LTDP, RDM, data science)
- Real world tests and benchmarking

CORE TRUST SEAL

FAIRSFAIR
Fostering Fair Data Practices in Europe

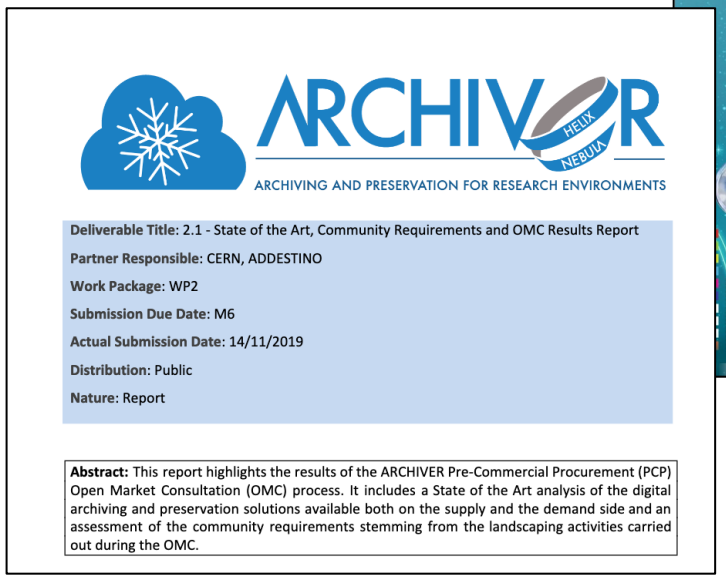
NDSA

CERN

PIC
port d'informació científica

EMBL-EBI

DESY


ARCHIVER
ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

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

Partner Responsible: CERN, ADDESTINO

Work Package: WP2

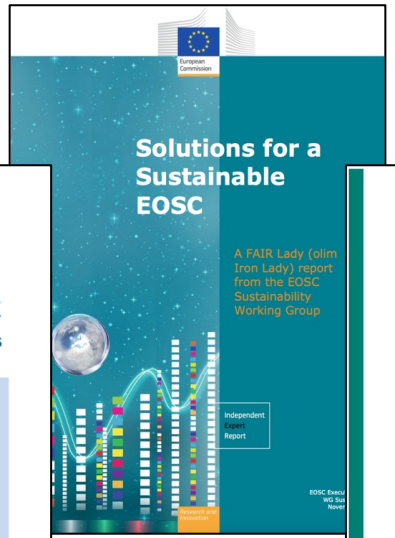
Submission Due Date: M6

Actual Submission Date: 14/11/2019

Distribution: Public

Nature: Report

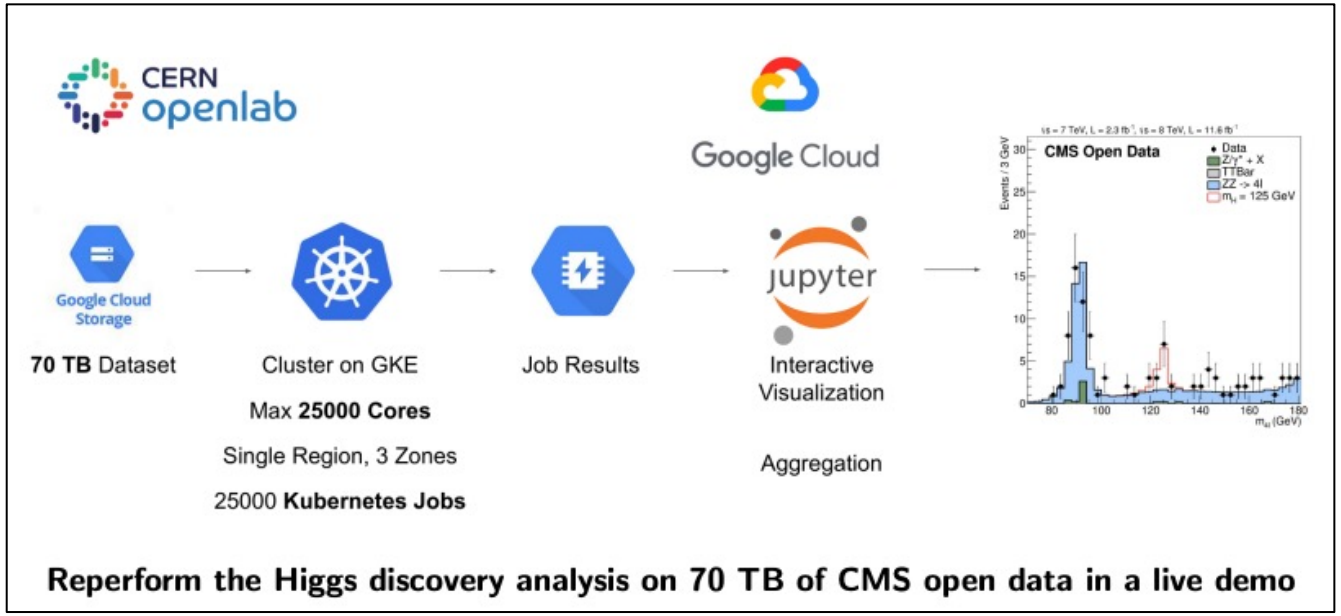
Abstract: This report highlights the results of the ARCHIVER Pre-Commercial Procurement (PCP) Open Market Consultation (OMC) process. It includes a State of the Art analysis of the digital archiving and preservation solutions available both on the supply and the demand side and an assessment of the community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



https://zenodo.org/record/4574234#.YW2NyqDTWJQ

Partnership with Google

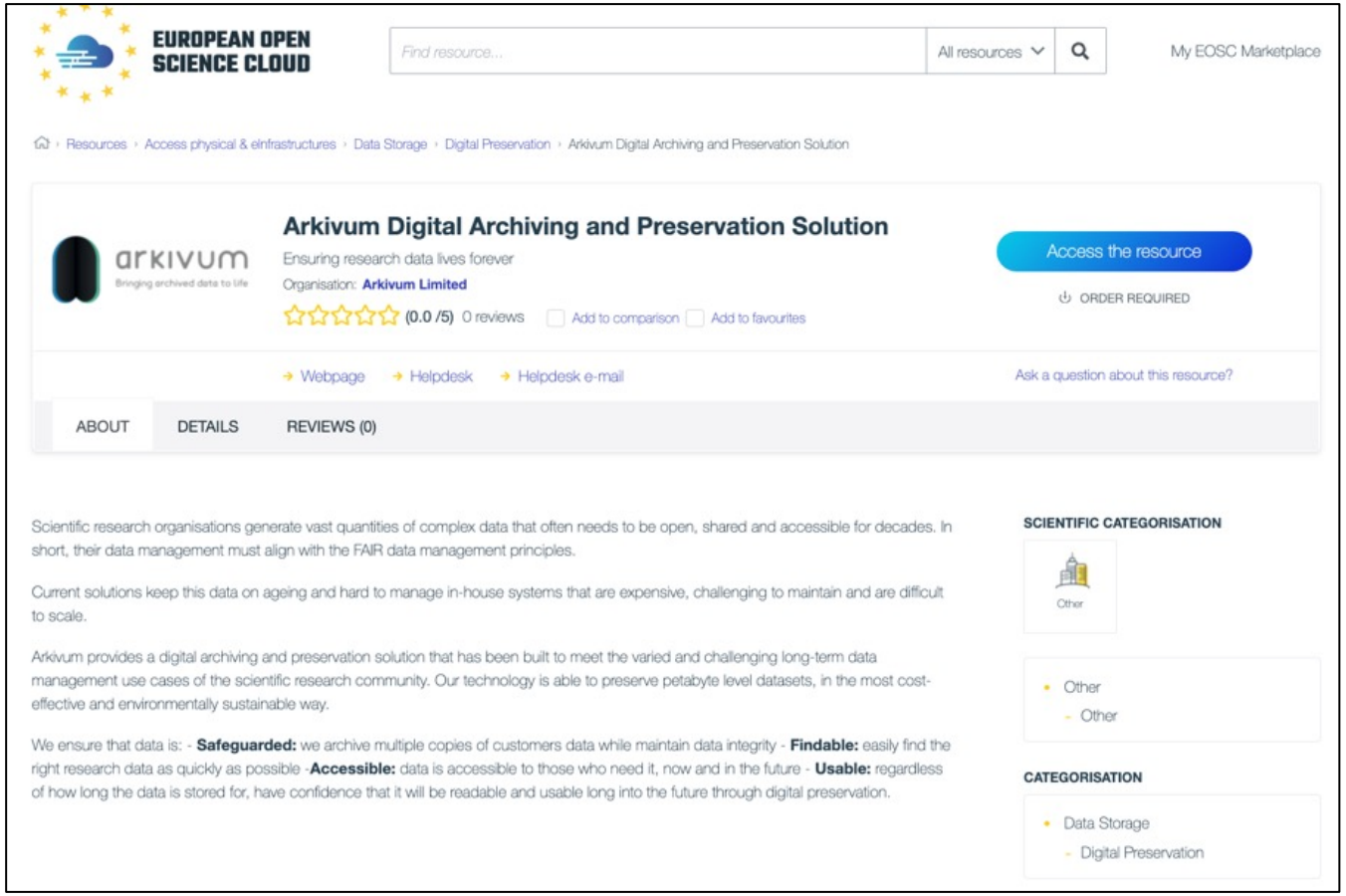
- GCP provides a highly scalable infrastructure (compute, storage, networking)
- Connected to GÉANT and NRENs across Europe
- GCP is widely used as a cloud platform for scientific data processing
- Low carbon footprint and good environmental credentials
- Discounts for education/research
- Arkivum is a Google Reseller
- Combined solution in ARCHIVER



https://indico.cern.ch/event/1008656/contributions/4232849/attachments/2192041/3705047/physics_analysis_google_meeting-16.02.2021.pdf

European Open Science Cloud (EOSC) and related projects

- EOSC Marketplace listing
  - Early Adopters Programme
- Ordering possible through EOSC
  - Via the Arkivum website
- Active involvement in past, present and future EOSC events
- Future initiatives and projects
  - EOSC TF DP

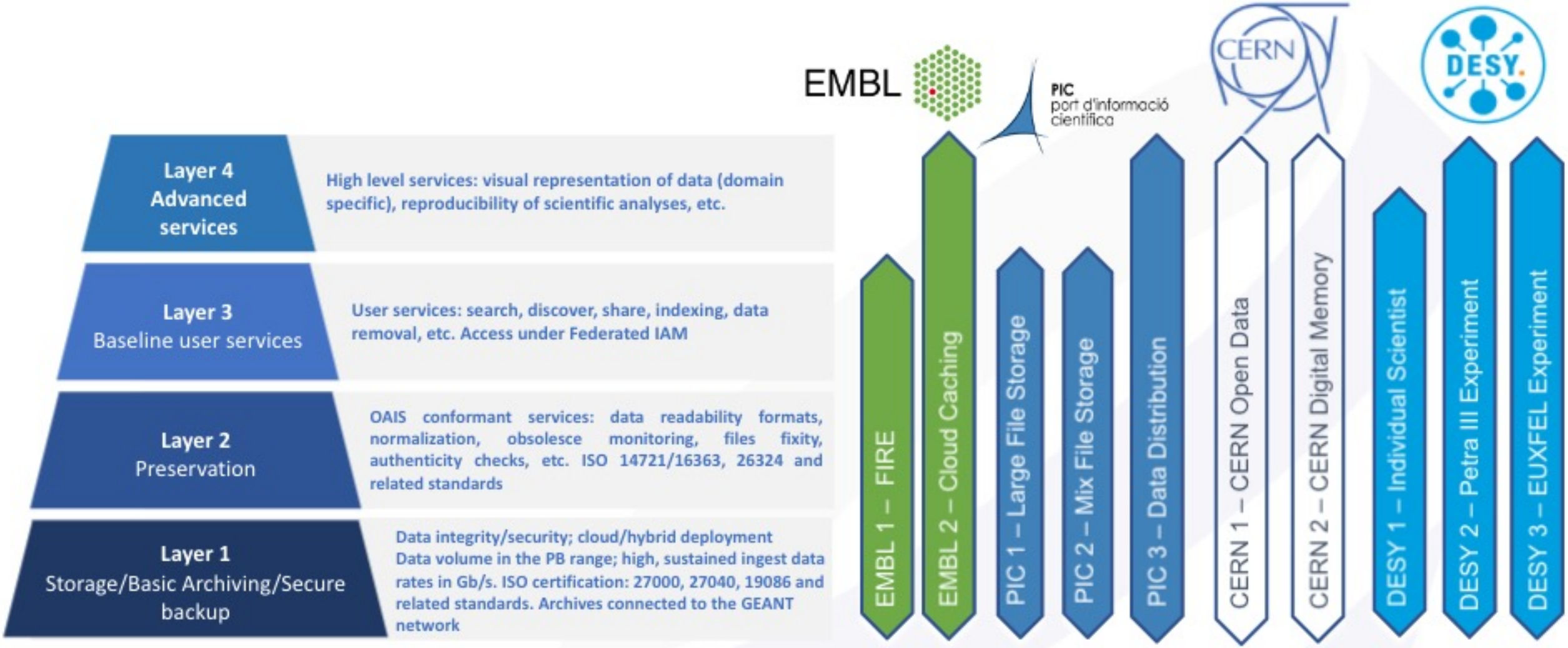


Solution Overview (20 mins)

Agenda

- LTDP of Research Datasets
- Architecture and Arkivum Solution
- Integration into RDM landscape of Repositories and FAIR
- SSO using eduGAIN
- Scalability and Performance
- Deployment options: cloud and on-premise
- Cost-efficiency and Environmental Sustainability

Long-Term Digital Preservation of Research Data: Functionality



Scientific use cases deployments: 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

Long-Term Digital Preservation of Research Data: Good Practice

TRUST

RDA
RESEARCH DATA ALLIANCE

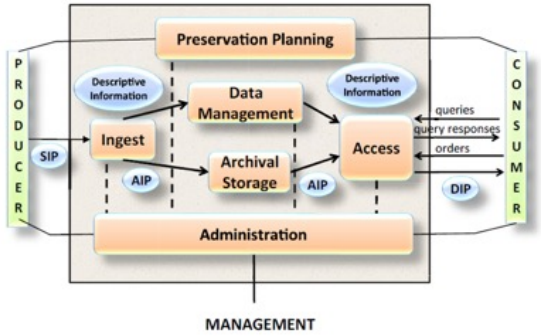
CORE TRUST SEAL

DigitalPreservationCoalition

FAIRSFAIR
Fostering Fair Data Practices in Europe

NDSA

Levels of Digital Preservation



F R13

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

A R10 R15 R16

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

I R14

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

R R2 R7 R14

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

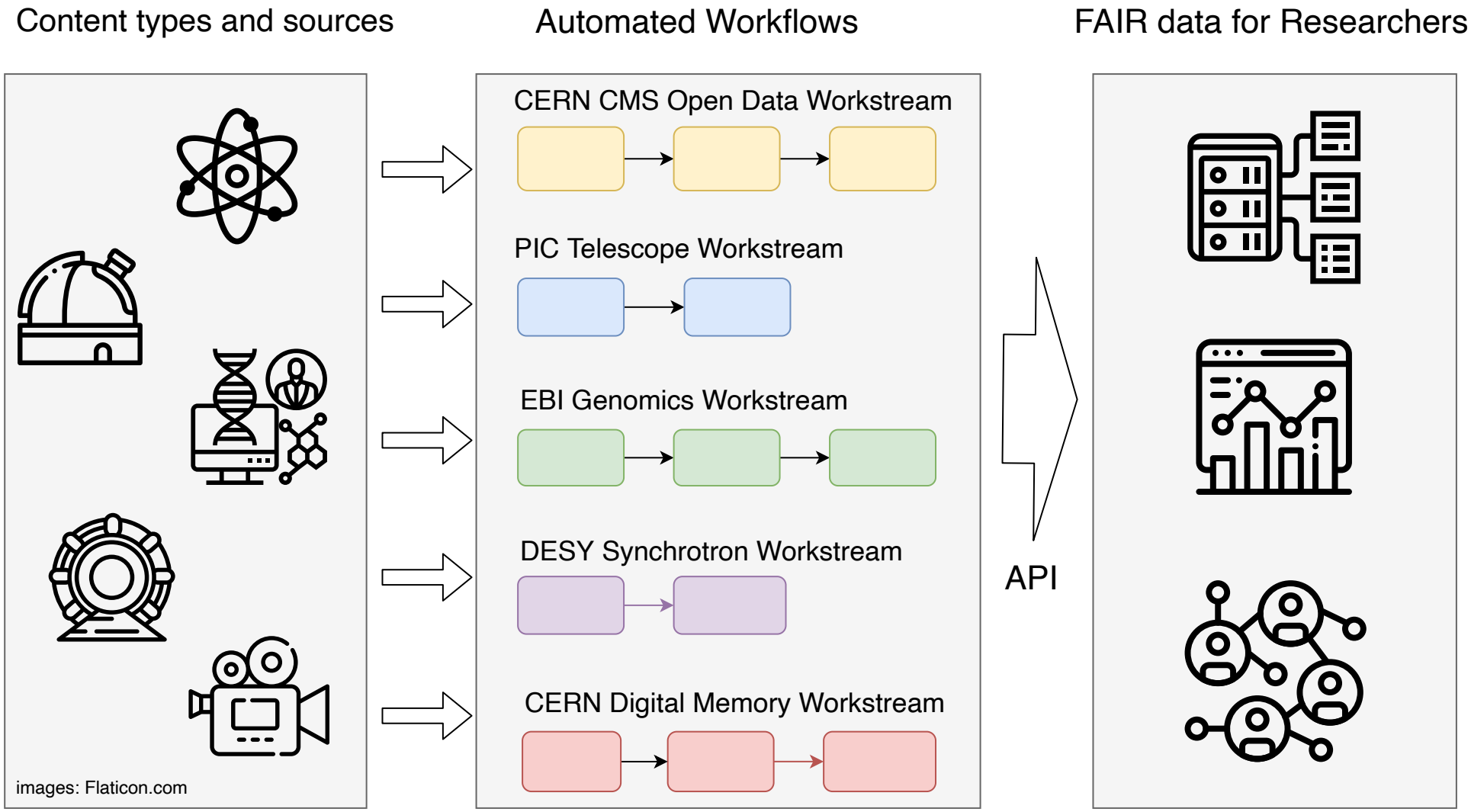
Standards, External Assessment and Certification

Self Assessment

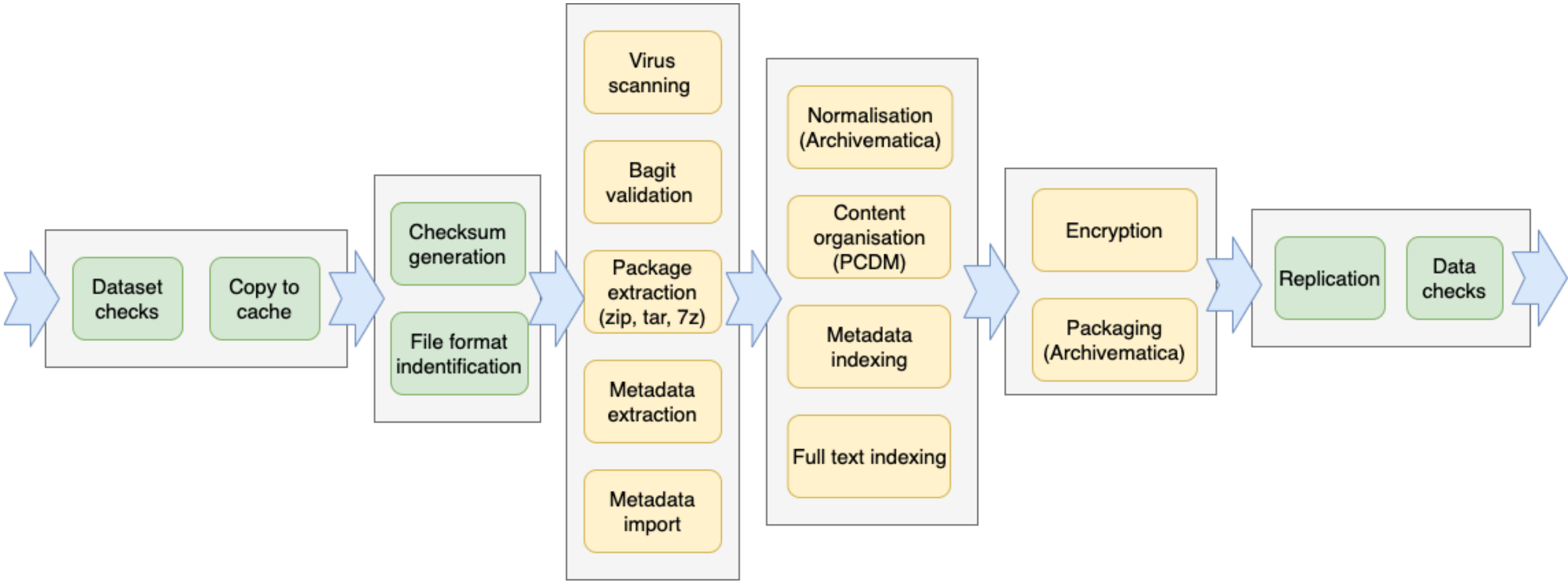
Peer Review

Formal Certification

Long-Term Digital Preservation (LTDP) Factories



Arkivum Microservices and Workflows



NDSA

Levels of Digital Preservation

NDSA Preservation Levels

Levels of Digital Preservation Assessment Tool

This tool can be used to assist you with determining which aspects of digital preservation you have strength in and which you may need to focus future efforts.

To use this tool, you will enter in a 0, 1, or 2 in each box next to a task. The conditional formatting will color code the tasks.

2 = Achieved

1 = Work in Progress

0 = Not started

NOTE: It is recognized that some of these tasks may not be 'in scope' for all organizations. For example, an organization may not be the ones in charge of a task and therefore entering a 0, 1, or 2 is not accurate. It is recommended to leave anything that your organization will not do or be able to do blank. This will keep the cell white allowing review to be done on only the colored cells.

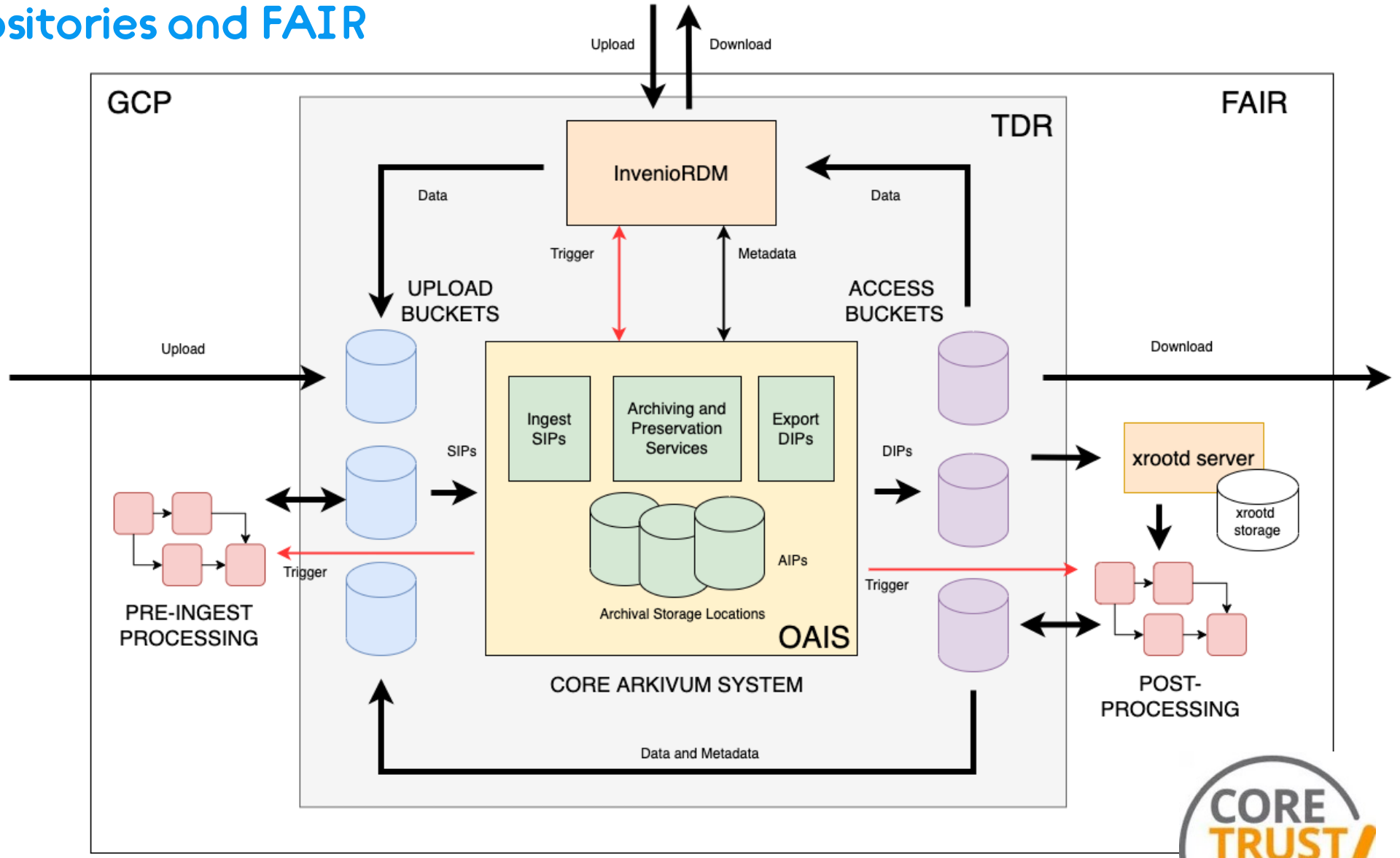
	Level							
Functional Area	ENTER 0, 1, 2	Level 1 (Know your content)	ENTER 0, 1, 2	Level 2 (Protect your content)	ENTER 0, 1, 2	Level 3 (Monitor your content)	ENTER 0, 1, 2	Level 4 (Sustain your content)
Storage	2	Have two complete copies in separate locations	2	Have three complete copies with at least one copy in a separate geographic location	2	Have at least one copy in a geographic location with a different disaster threat than the other copies	2	Have at least three copies in geographic locations, each with a different disaster threat
	2	Document all storage media where content is stored	2	Document storage and storage media indicating the resources and dependencies they require to function	2	Have at least one copy on a different storage media type	2	Maximize storage diversification to avoid single points of failure
	2	Put content into a stable storage			2	Track the obsolescence of storage and media	2	Have a plan and execute actions to address obsolescence of storage hardware, software, and media
Integrity	2	Verify integrity information if it has been provided with the content	2	Verify integrity information when moving or copying content	2	Verify integrity information of content at fixed intervals	2	Verify integrity information in response to specific events or activities
	2	Generate integrity information if not provided with the content		Use write-blockers when worki original media				
	2	Virus check all content; isolate content for quarantine as needed	2	Back up integrity information an copy in a separate location fron content				
Control	2	Determine the human and software agents that should be authorized to read, write, move, and delete content	2	Document the human and softw agents authorized to read, write and delete content and apply th				
Metadata	2	Create inventory of content, also documenting current storage locations	2	Store enough metadata to know content is (this might include so combination of administrative, te descriptive, preservation, and s				
	2	Back up inventory and store at least one copy separately from content						
Content	2	Document file formats and other essential content characteristics including how and when these were identified	2	Verify file formats and other ess content characistics				
				Build relationships with content to encourage sustainable file ch				

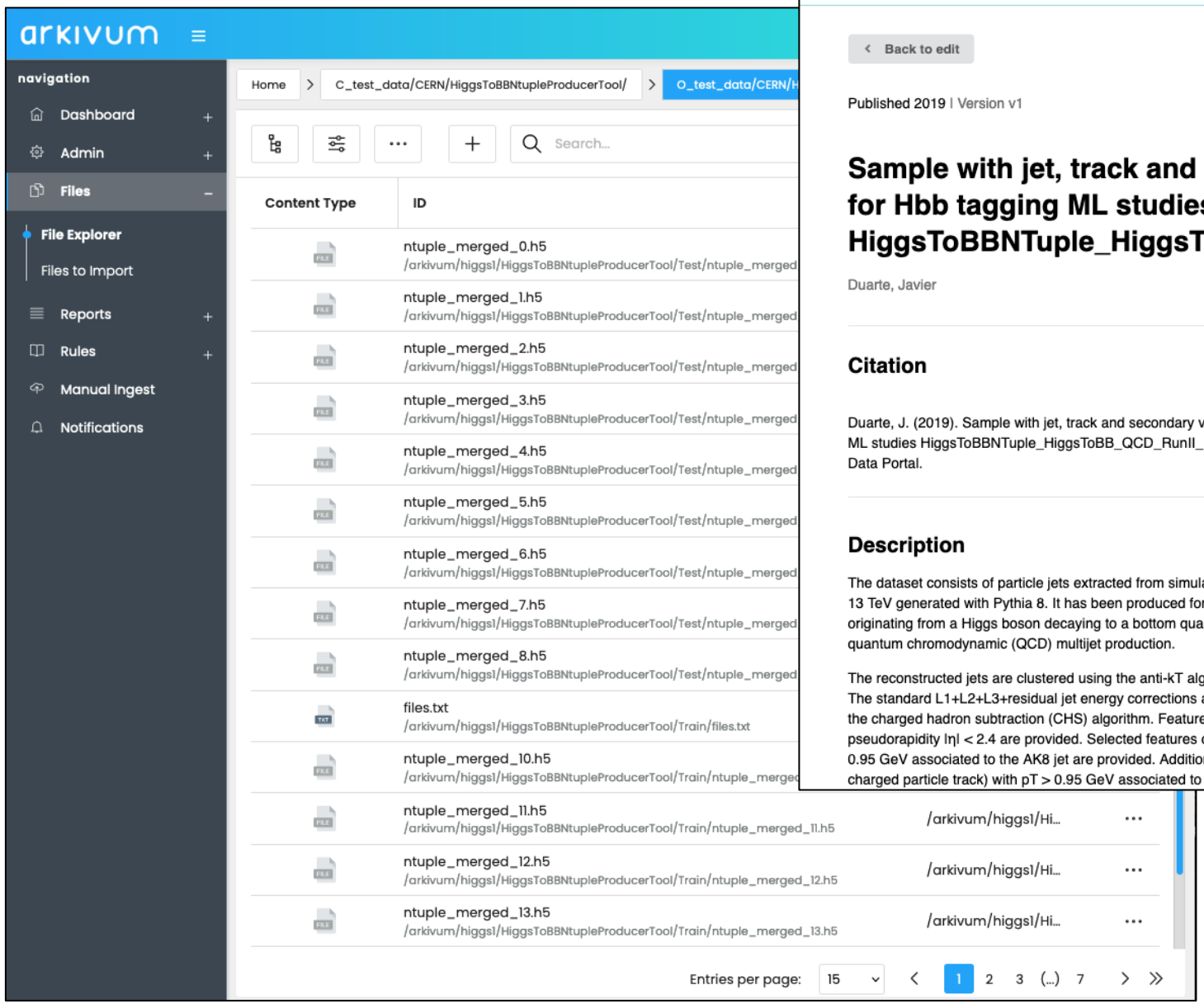
Integrity	
	• Content can be ingested using customer provided checksums that are used to verify data integrity at the point of ingest (e.g. using bagit). • Content can be exported with checksums and the content is validated at the point of export (e.g. supplied as validated bagit bags). • Checksums are generated for any content ingested that doesn't have a customer supplied checksum. • Checksums are used to validate data integrity when content is replicated and stored in different storage locations. This includes when content is first stored and also periodically (e.g. annual checks). • Checksums are used to validate data integrity when archive content is retrieved, e.g. to check the fixity of a file when it is downloaded by a user. • Content is virus checked as part of the ingest process. • Multiple copies of content (see Storage above) along with checksum validation are used to repair data corruptions or loss. • All integrity related actions (checksum validation on ingest, checksum generation, periodic fixity checks etc.) are all recorded in an immutable and protected system audit trail. • Integrity reports are available to show customers when fixity was last checked and the outcome of the check. This allows customers to audit integrity and fixity checking. • Checksums and fixity checks are done in addition to, and independently of, whatever checks are done by the underlying storage location. For example, when storing data in AWS/Azure/Google object storage, the solution will verify that the data has been written correctly and will do this independently of whatever assertions or assurances are provided by the storage layer/storage provider.

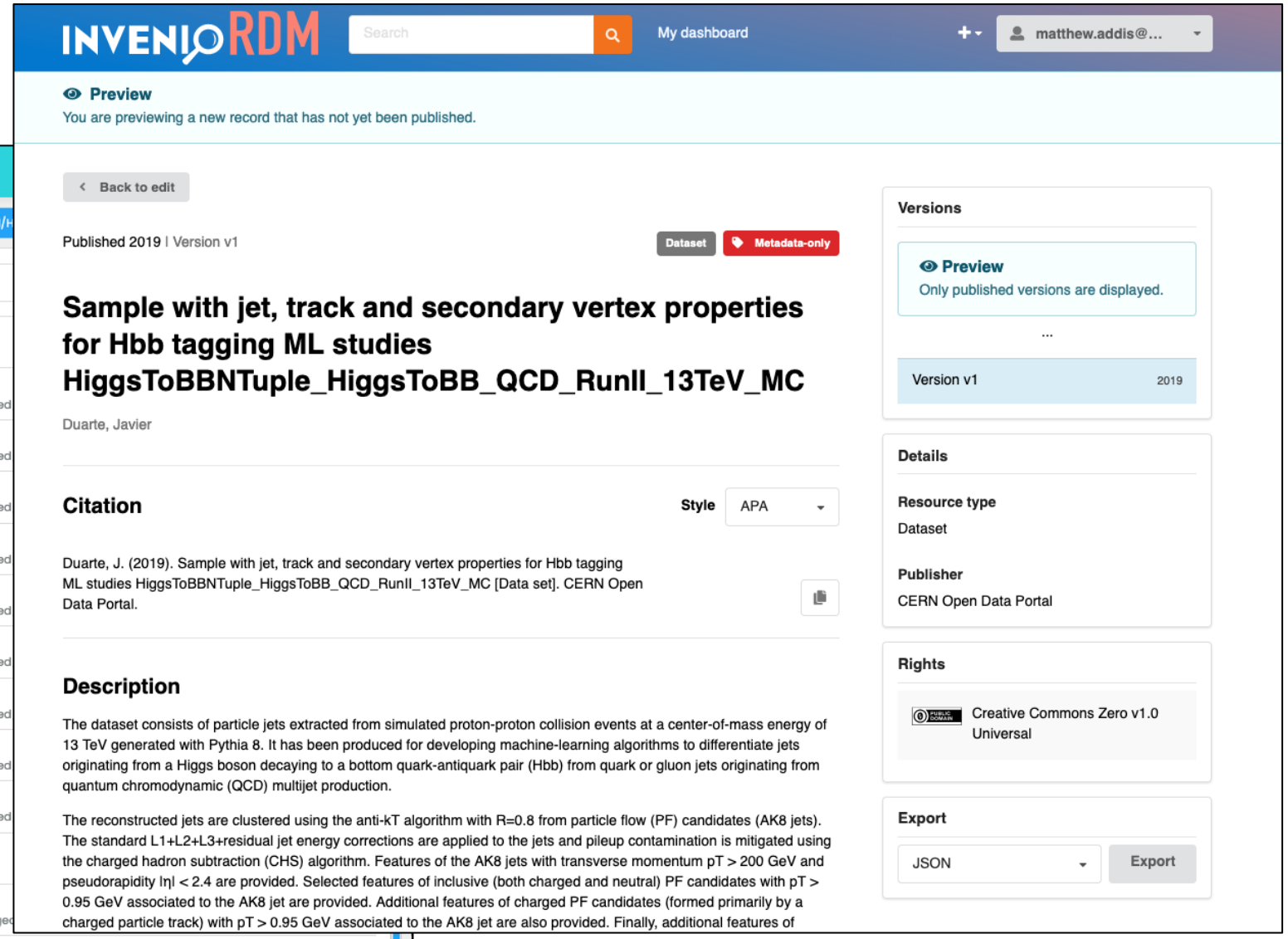
DPC RAM

I - Content preservation Processes to preserve the meaning or functionality of the digital content and ensure its continued accessibility and usability over time	3 - Managed	The organization has implemented a managed process to monitor and plan for accessibility of content over time, for example: • Technology watch activities are carried out and 'at risk' content is identified. • Technical dependencies are detected and documented. • Actions are occasionally carried out to ensure preservation and quality of content such as migration, emulation or modification of creation or capture workflows. • Preservation actions occur with an understanding of the properties of the digital object that should be retained to support current and future use cases. • All changes to digital content are recorded, including details of when, what, how, why and who.	• The solution supports file format identification using a range of tools (Tika, Siegfried, FIDO). • The system records a wide range of file characteristics including file type, date, size, filepath and checksums. • Technical metadata extraction and file format characterisation provides detailed technical information on the types of content being stored in the solution. • File format verification is supported, e.g. using JHOVE, MediaConch and VeraPDF. • File format migrations/normalisations can be done according to configurable rules, e.g. using the Archivematica Format Policy Register. • Users can search for content according to file format or other technical characteristics. This is in addition to searching by descriptive metadata. • Reports are available on content in the solution, e.g. normalisation reports on file format conversion done during ingest. • File format identification, characterisation, validation and normalisation are all recorded in the system audit trail.	The Arkivum solution supports the execution of preservation actions, e.g. file format normalisations, but it is up to the organisation using the solution to assess their risks and decide the preservation plans/policies most suitable for them.

Repositories and FAIR




arkivum
navigation
Dashboard
Admin
Files
File Explorer
Files to Import
Reports
Rules
Manual Ingest
Notifications
Home
C_test_data/CERN/HiggsToBBNtupleProducerTool/
O_test_data/CERN/H
Search...
Content Type
ID
ntuple_merged_0.h5
/arkivum/higgs1/HiggsToBBNtupleProducerTool/Test/ntuple_merged
ntuple_merged_1.h5
/arkivum/higgs1/HiggsToBBNtupleProducerTool/Test/ntuple_merged
ntuple_merged_2.h5
/arkivum/higgs1/HiggsToBBNtupleProducerTool/Test/ntuple_merged
ntuple_merged_3.h5
/arkivum/higgs1/HiggsToBBNtupleProducerTool/Test/ntuple_merged
ntuple_merged_4.h5
/arkivum/higgs1/HiggsToBBNtupleProducerTool/Test/ntuple_merged
ntuple_merged_5.h5
/arkivum/higgs1/HiggsToBBNtupleProducerTool/Test/ntuple_merged
ntuple_merged_6.h5
/arkivum/higgs1/HiggsToBBNtupleProducerTool/Test/ntuple_merged
ntuple_merged_7.h5
/arkivum/higgs1/HiggsToBBNtupleProducerTool/Test/ntuple_merged
ntuple_merged_8.h5
/arkivum/higgs1/HiggsToBBNtupleProducerTool/Test/ntuple_merged
files.txt
/arkivum/higgs1/HiggsToBBNtupleProducerTool/Train/files.txt
ntuple_merged_10.h5
/arkivum/higgs1/HiggsToBBNtupleProducerTool/Train/ntuple_merged
ntuple_merged_11.h5
/arkivum/higgs1/HiggsToBBNtupleProducerTool/Train/ntuple_merged_11.h5
/arkivum/higgs1/Hi...
ntuple_merged_12.h5
/arkivum/higgs1/HiggsToBBNtupleProducerTool/Train/ntuple_merged_12.h5
/arkivum/higgs1/Hi...
ntuple_merged_13.h5
/arkivum/higgs1/HiggsToBBNtupleProducerTool/Train/ntuple_merged_13.h5
/arkivum/higgs1/Hi...
Entries per page:
15
1
2
3
(...)
7

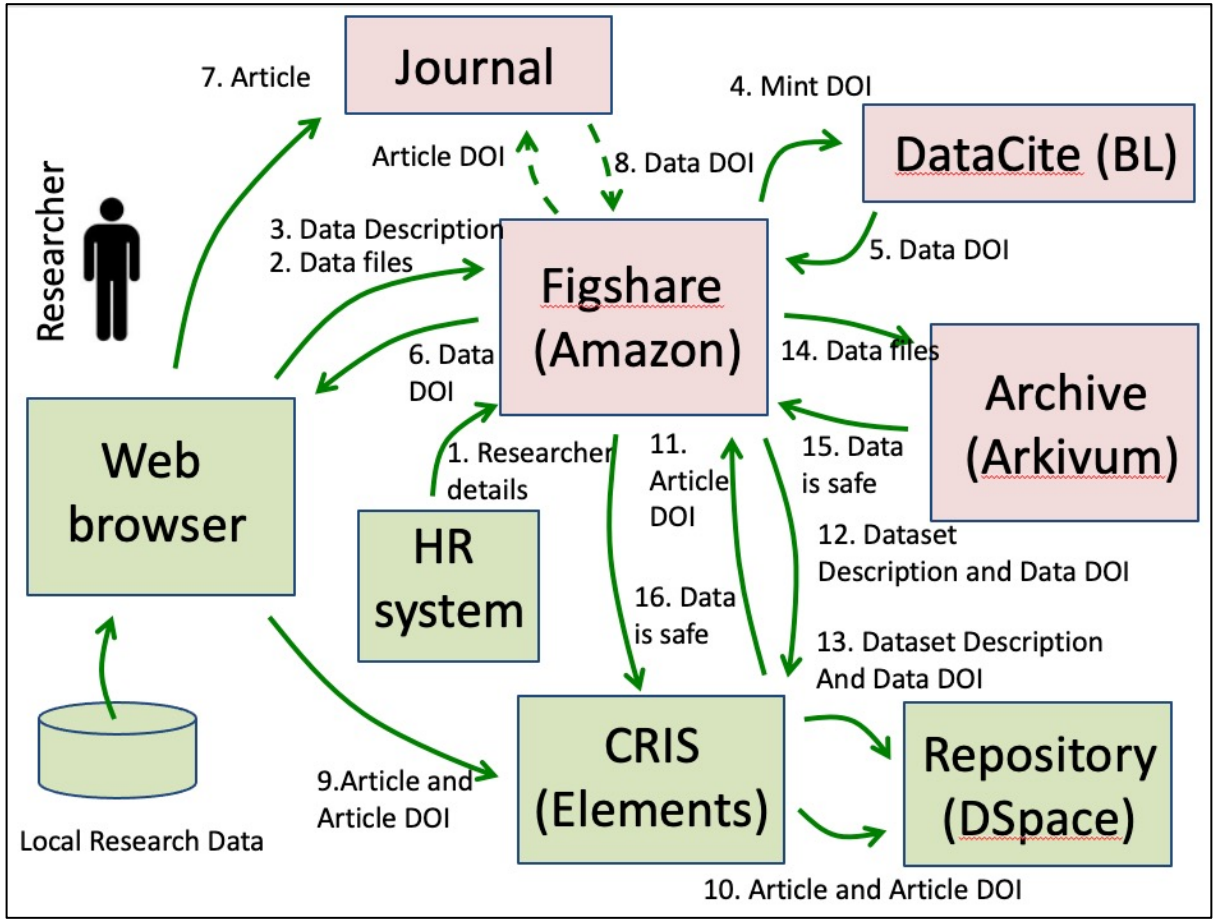

INVENIO RDM
Search
My dashboard
matthew.addis@...
Preview
You are previewing a new record that has not yet been published.
Back to edit
Published 2019 | Version v1
Dataset
Metadata-only
Sample with jet, track and secondary vertex properties for Hbb tagging ML studies HiggsToBBNTuple_HiggsToBB_QCD_RunII_13TeV_MC
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
JSON
Export

Other Repository Integrations

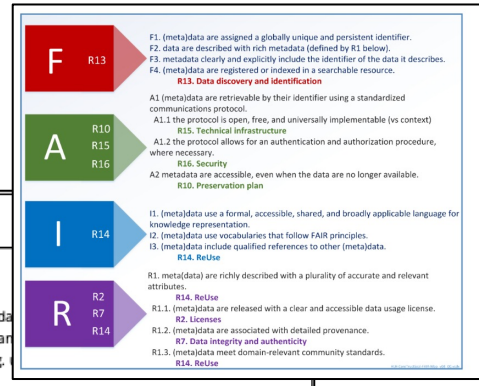
figshare

DSPACE

eprints
repository software



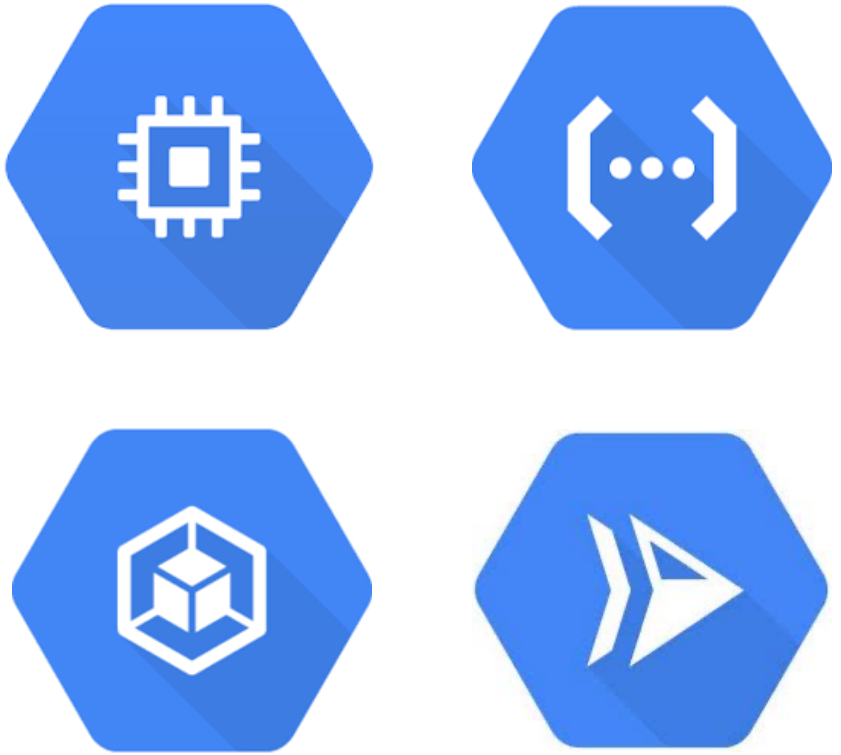
CoreTrustSeal and FAIR



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. Metada Metadata can be searched and retrieved using a REST API. Metadata files can including json-ld, and these metadata files can be associated to datasets, e.g. (Metadata in InvenioRDM is in json format and can be searched and exported via a REST API. InvenioRDM landing pages include record citation metadata in machine readable form (json).
		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 format (XML or json). These metadata files can include references to external schemas or vocabs, e.g. schema.org, wikidata. 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.

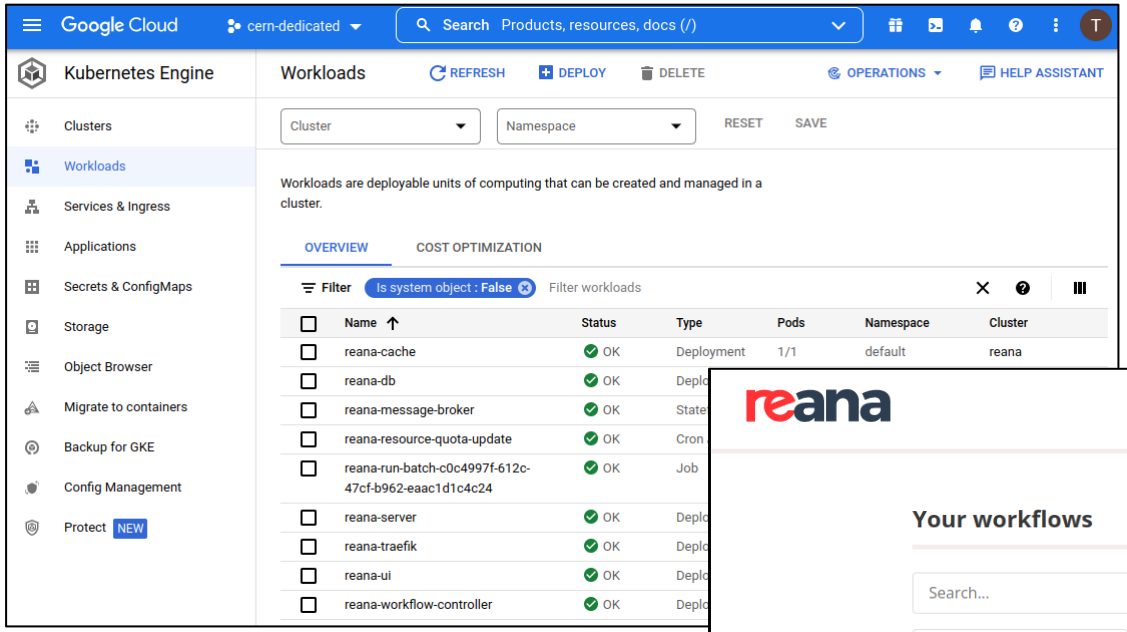
Cloud-hosted Processing of Archived Research Datasets

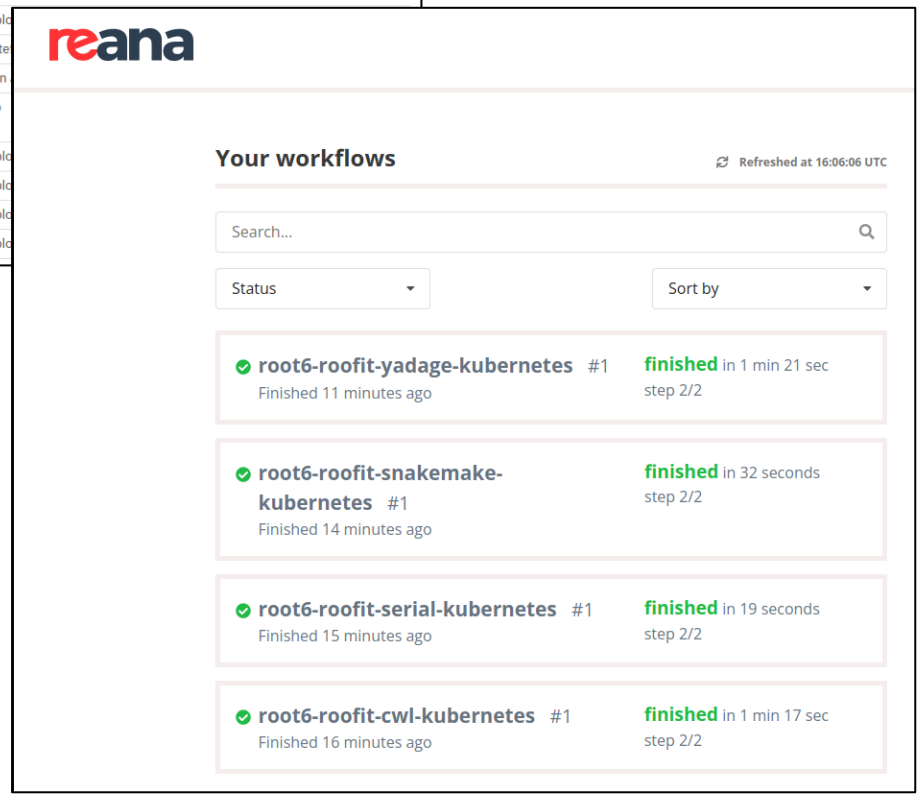
- GCP hosted applications
  - Cloud Functions (serverless code)
  - Cloud Run (containers)
  - GKE (Kubernetes)
  - Compute Engine (VMs)
- Arkivum Webhooks
  - Ingest, preservation, access
- Arkivum REST API
  - including search, get metadata, export files
- Xrootd server
  - Easy integration of scientific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

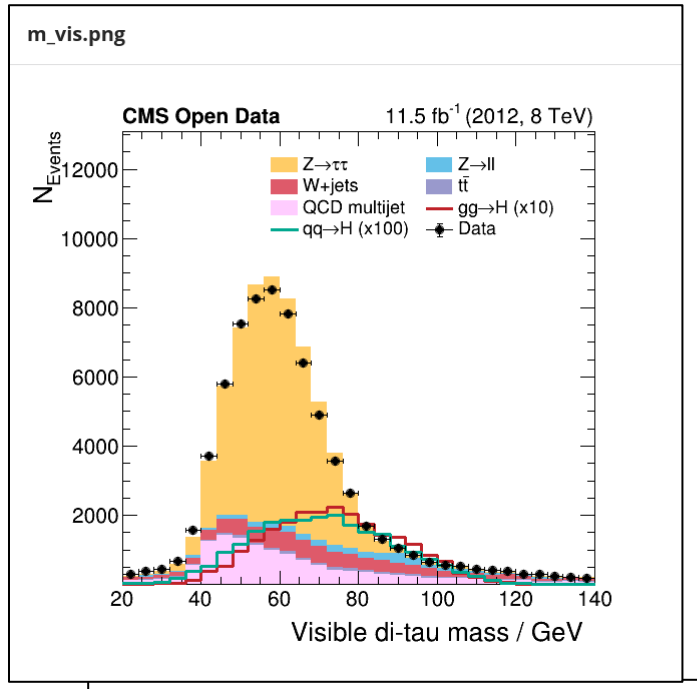


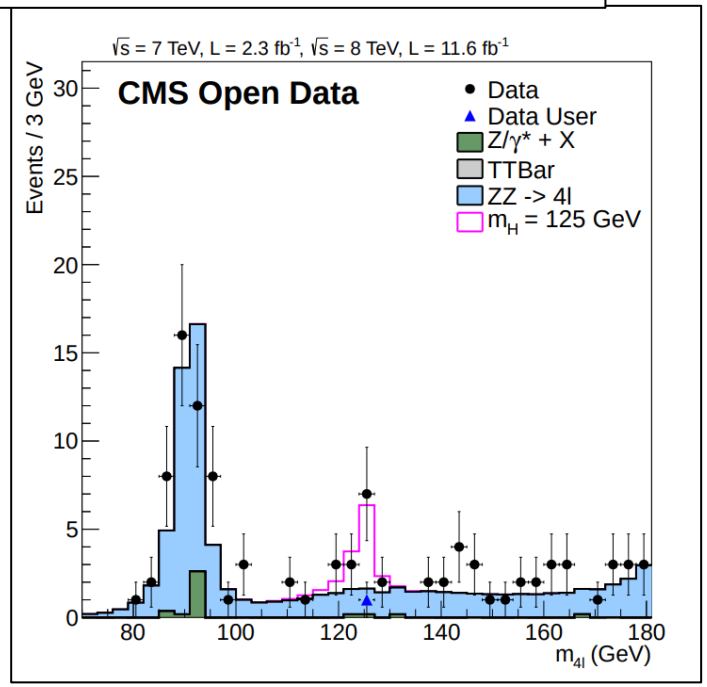
reana

snakemake









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)

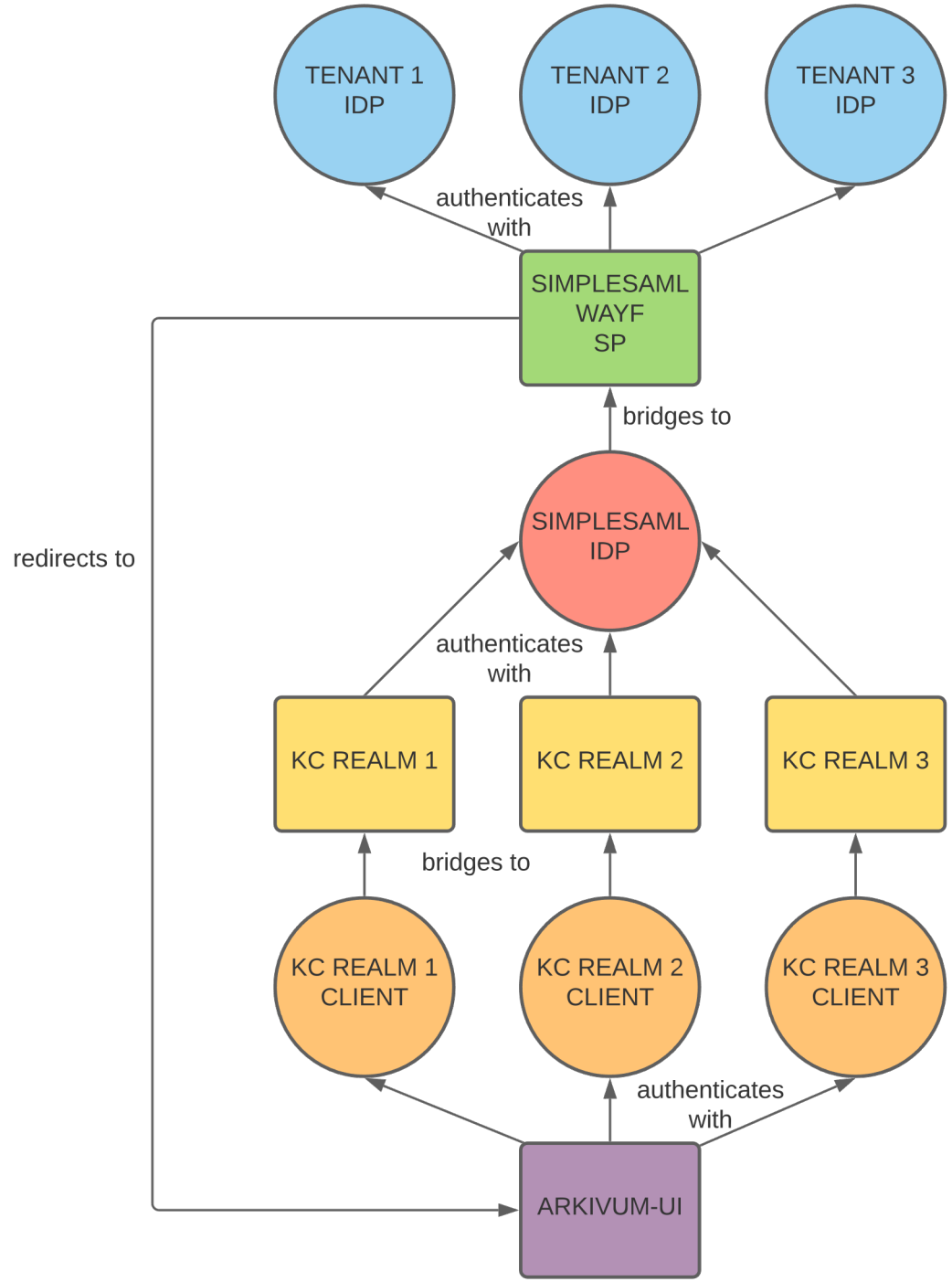
Open Standards, Open Specifications, Open Source: Deployment

Open Standards, Open Specifications, Open Source: Interaction

Bagit & BDbags

eduGAIN and AAI

- Arkivum is a member of the UKAMF
- eduGAIN integration uses SimpleSAML php
  - WAYF service for selecting IdP
  - Routing of IdP authentication to systems/tenants
  - Mapping of user's IdP attributes to roles/groups
- Other SSO options possible:
  - LDAP and Active Directory
  - SAML and OpenID
  - Social (Twitter, Facebook, Instagram, Google etc.)

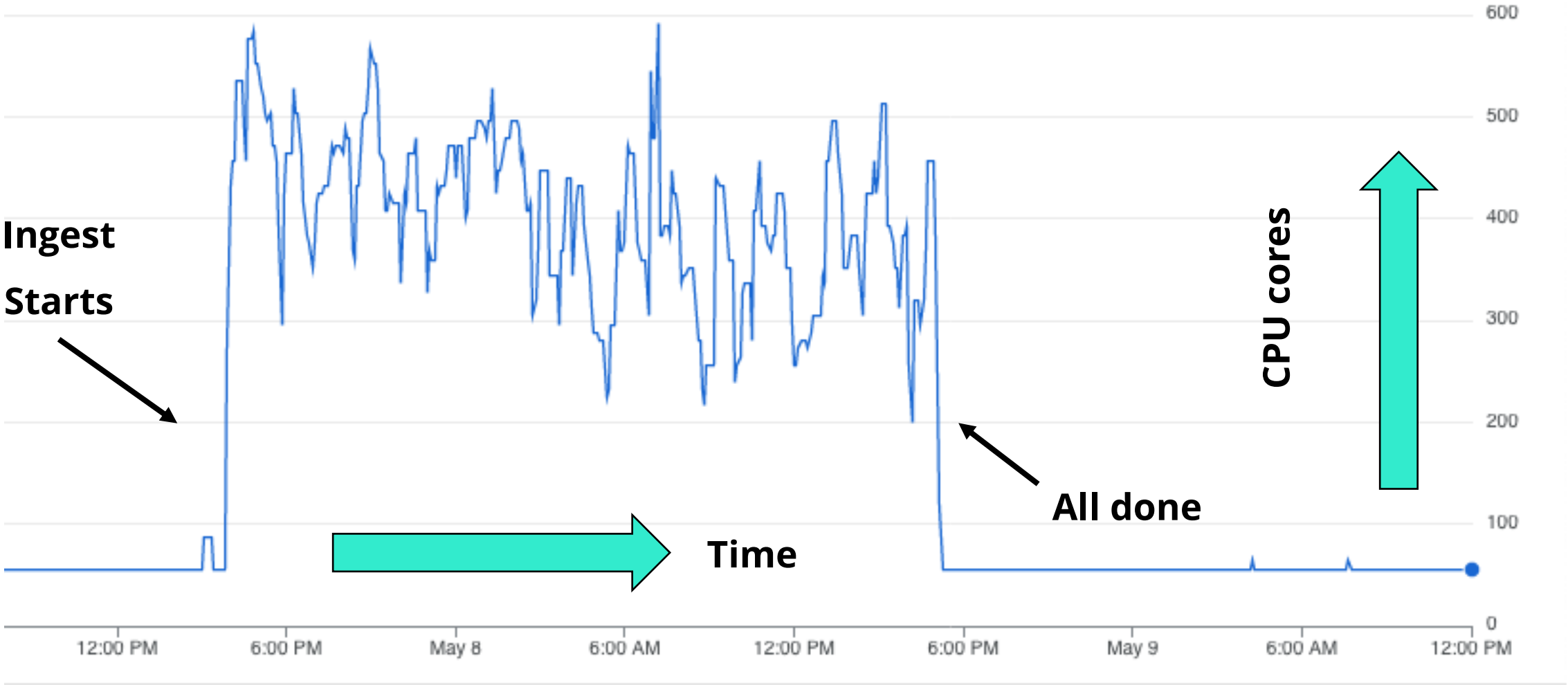


Scalability and Performance

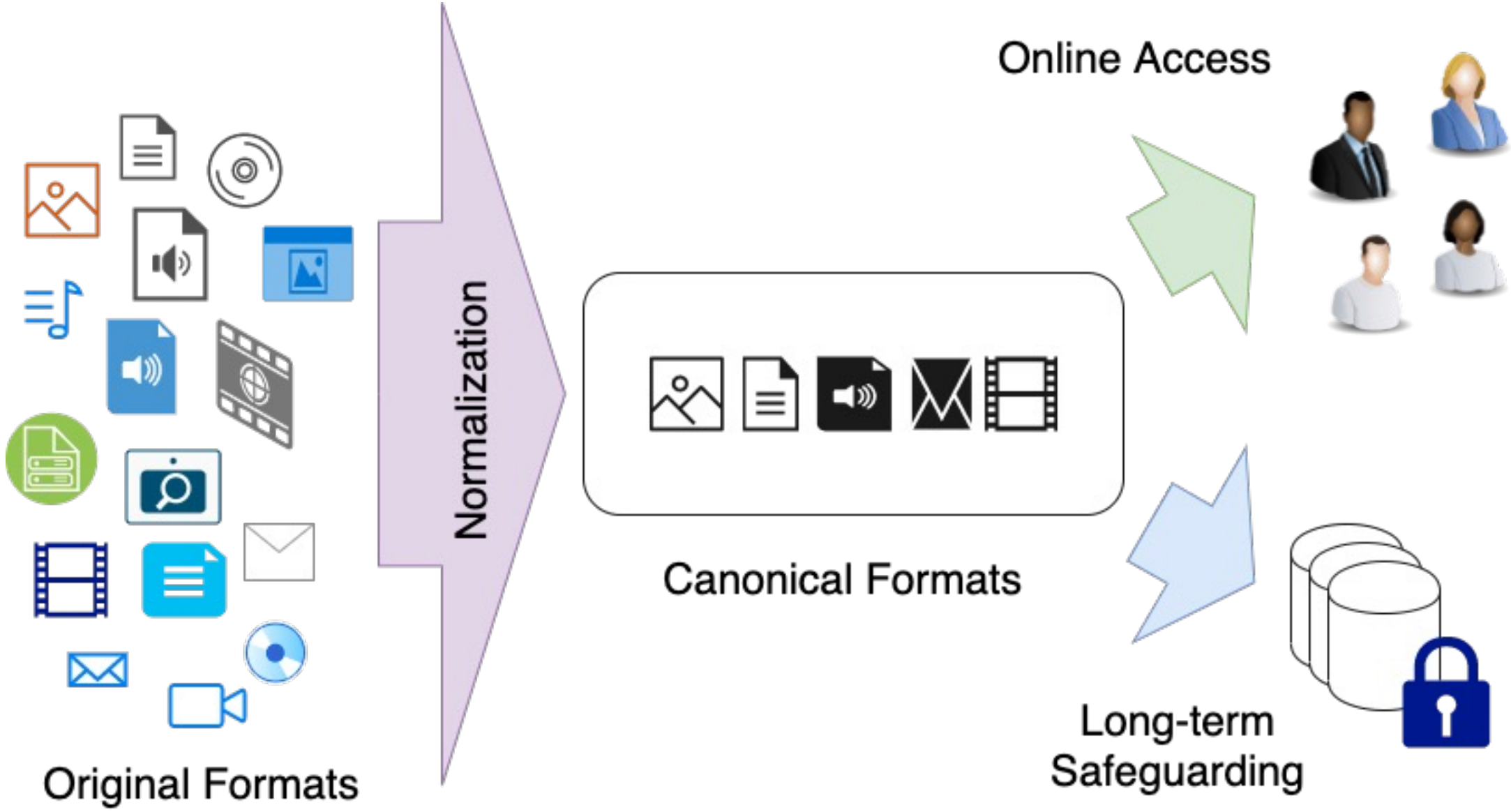
Serverless Computing: Scalability, Performance, Efficiency

- 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
Original Formats
Long-term
Safeguarding

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

Preferred Formats

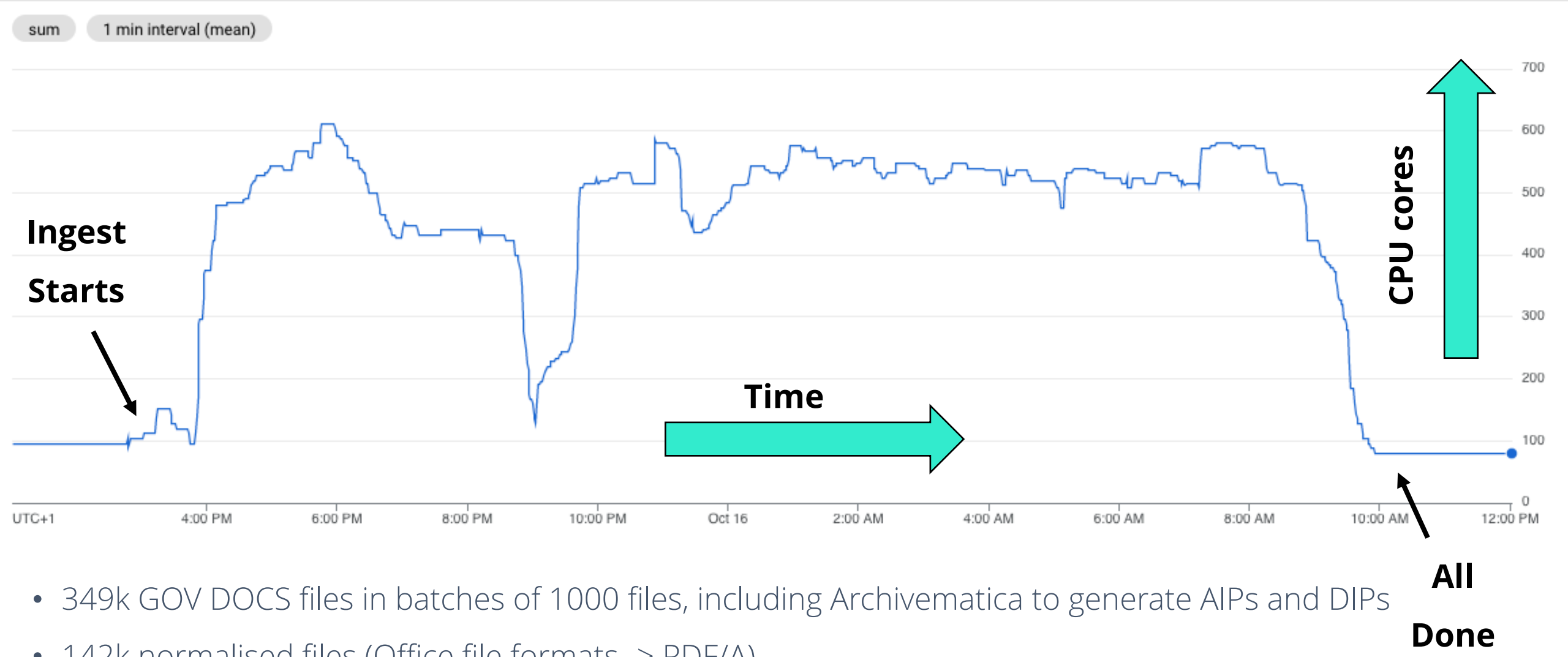
Preferred Formats	Format Versions	Format Specifications
ASCII Text	7 bit	ISO/IEC 646:1991 Information technology -- ISO 7-bit coded character set for information interchange: (http://www.iso.org/iso/catalogue_detail.htm?csnumber=4777)
Unicode Text	UTF-8	RTF 3629: UTF-8, A Transformation Format of ISO 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)


Cloud Deployment, Costs, Environmental Sustainability

Leveraging the Scalability and Efficiency of Cloud Infrastructure

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[3], 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

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

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*).

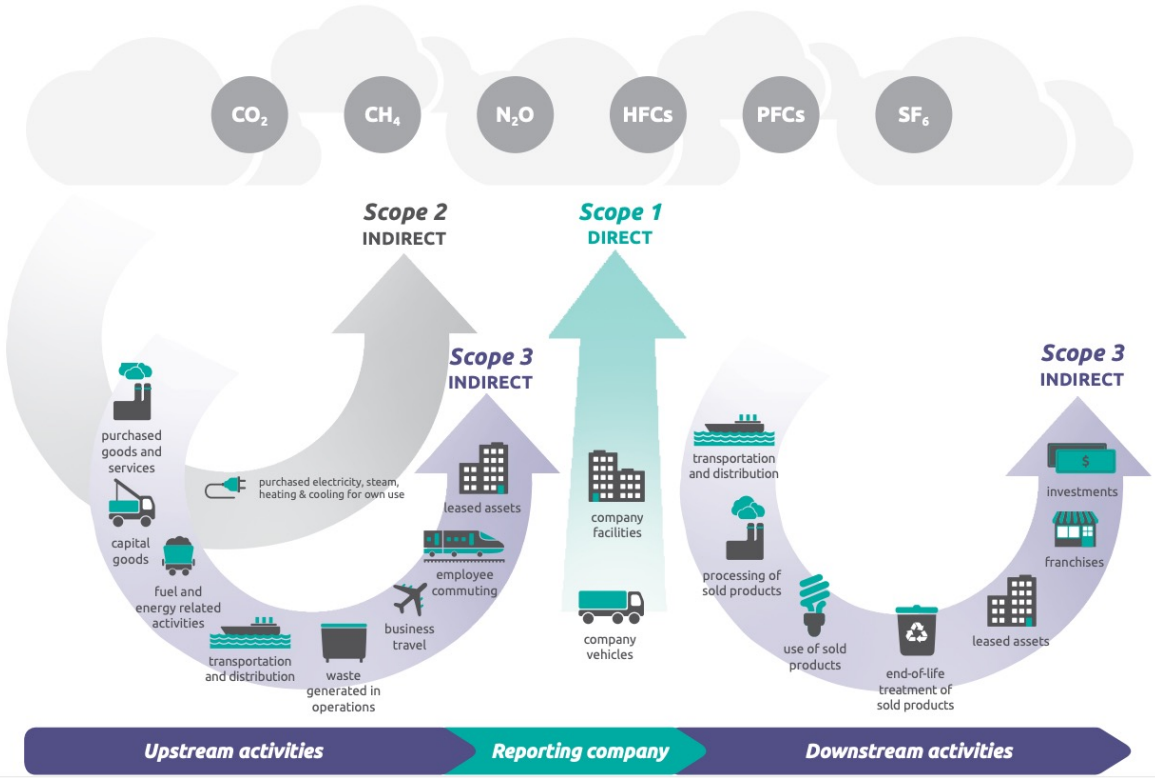
Since 2010, however, the data center landscape has changed dramatically (see the first figure). By 2018, global data center workloads and compute instances had increased more than sixfold, whereas data center internet protocol (IP) traffic had increased by more than 10-fold (*1*). Data center storage 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).

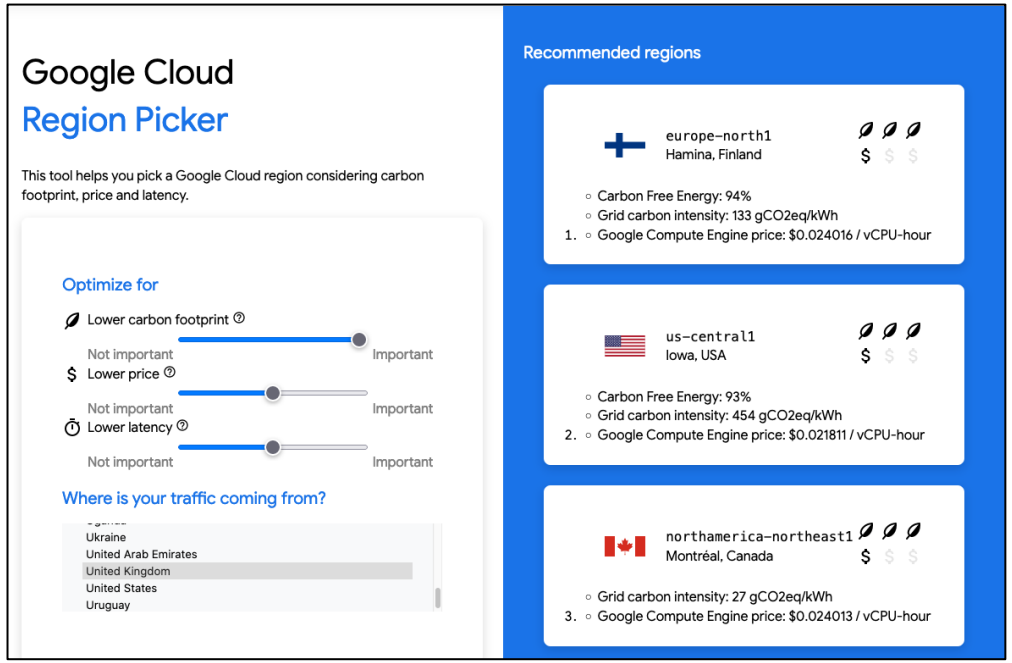
But since 2010, electricity use per computation of a typical volume server—the workhorse of the data center—has dropped by a factor of four, largely owing to processor-efficiency improvements and reductions in idle power (*10*). At the same time, the watts per terabyte of installed storage has dropped 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

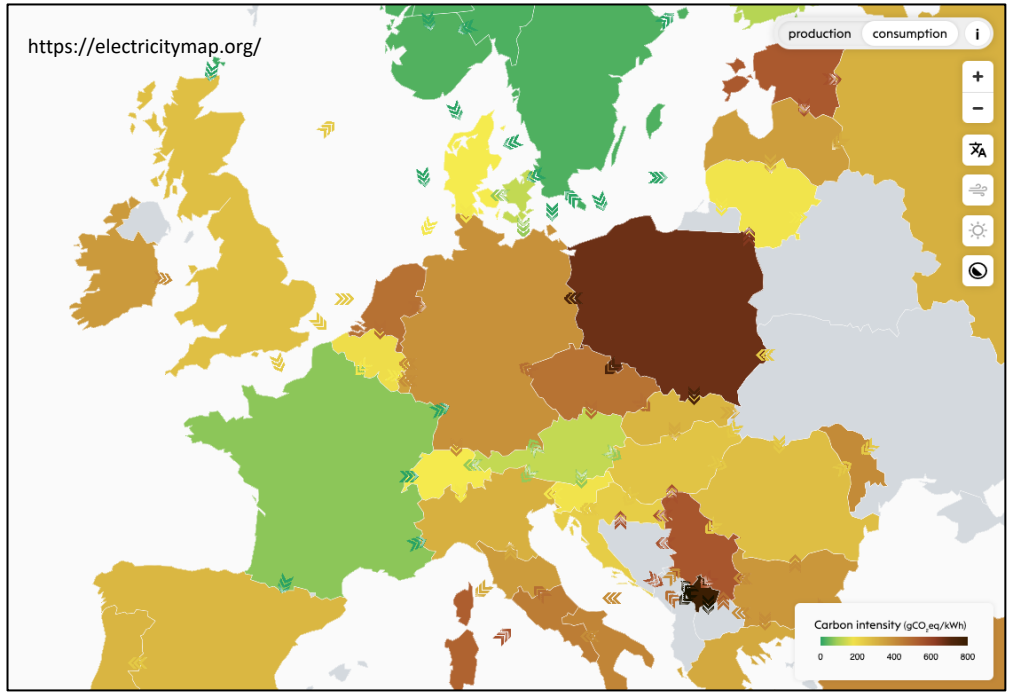
https://datacenters.lbl.gov/sites/default/files/Masanet_et_al_Science_2020.full_.pdf

Green Data Centers, Renewable Energy

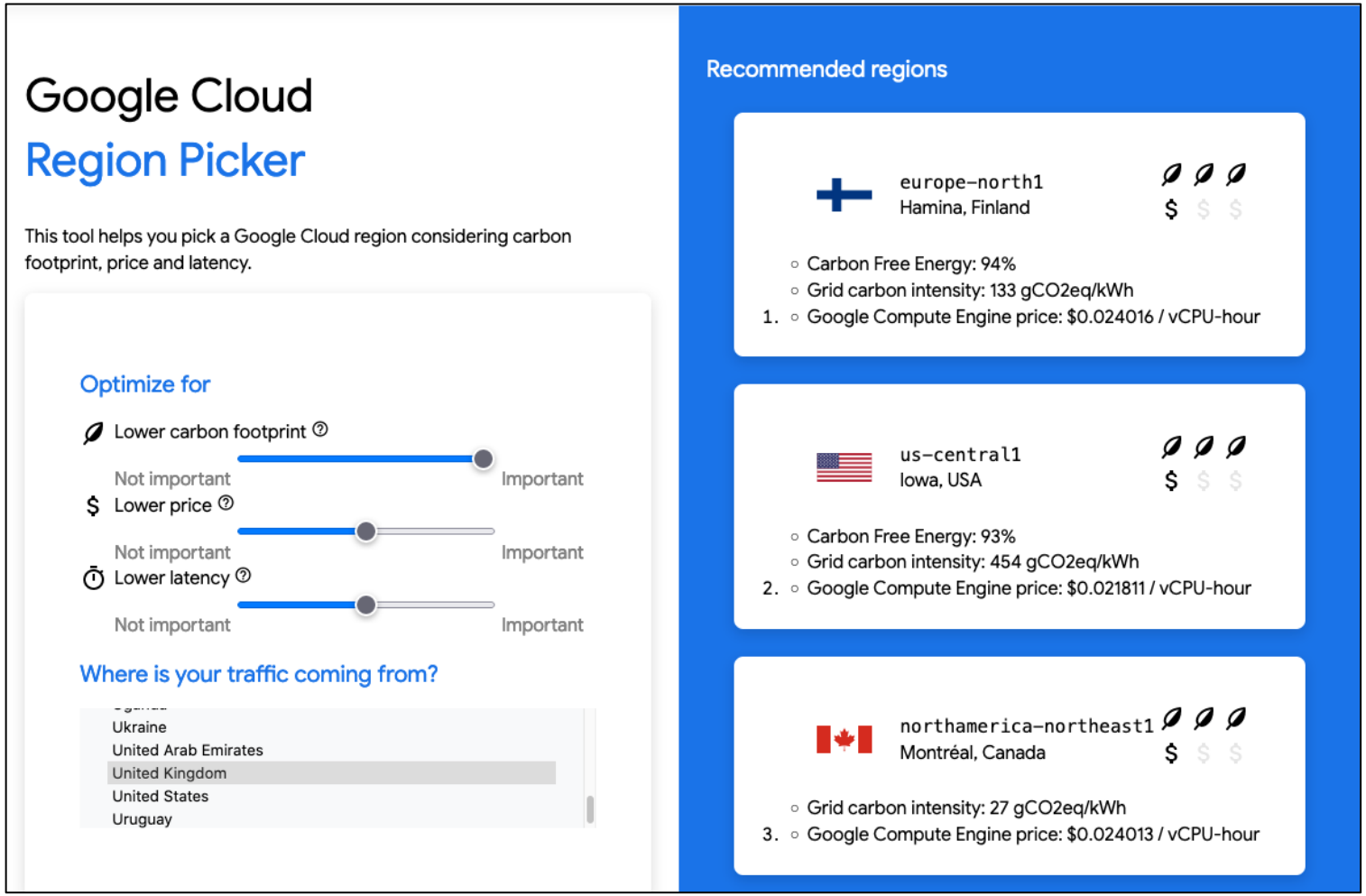
Figure [1.1] Overview of GHG Protocol scopes and emissions across the value chain







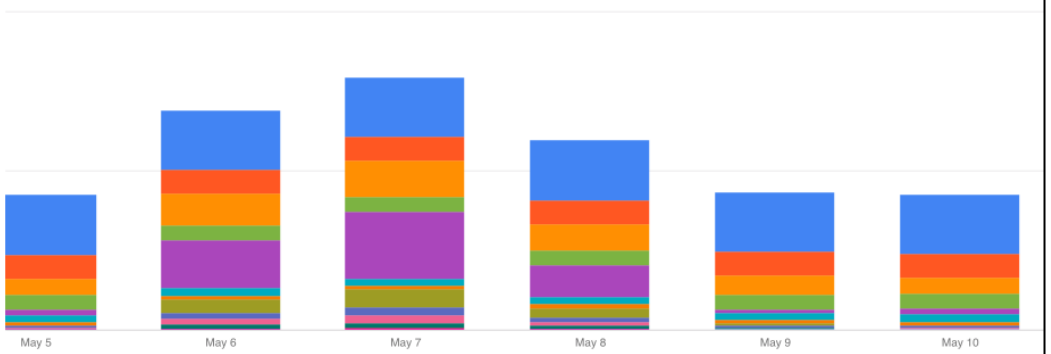
Green Data Centers, Renewable Energy



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

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

Costs, Resource Consumption, Carbon Footprint



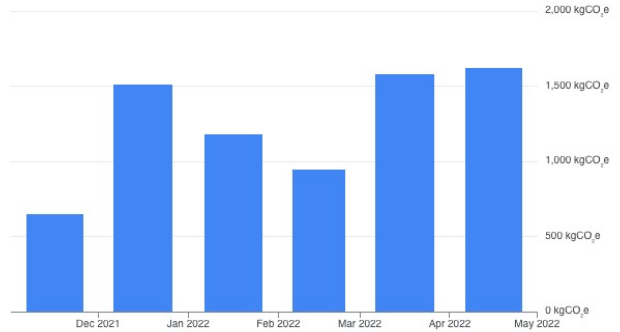
SKU	Service	SKU ID	Usage	Cost	Discounts
Custom Instance Core running in Frankfurt	Compute Engine	47BE-44D0-C86F	6,720 hour	£195.89	–
Custom Instance Ram running in Frankfurt	Compute Engine	6971-C36E-2B52	20,160 gibibyte hour	£78.75	–
Spot Preemptible N2D AMD Instance Core running in Frankfurt	Compute 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 $kgCO_2e$
From November 2021 to April 2022

April 2022 gross carbon footprint
1,621 $kgCO_2e$
↑ 2.78% comparing to March 2022

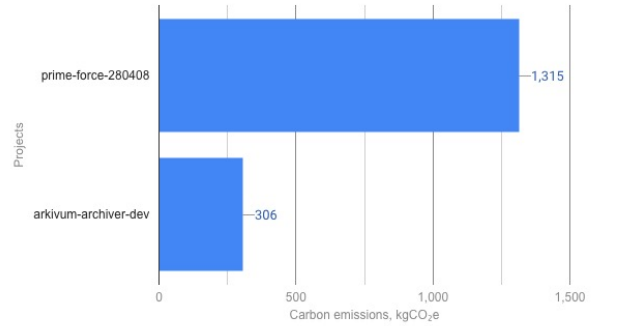
Google Cloud's net operational greenhouse gas emissions
0 $kgCO_2e$

Gross monthly carbon emissions



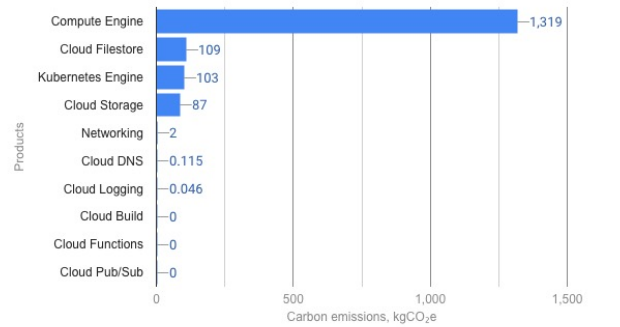
Gross carbon emissions by project in April 2022

Chart view



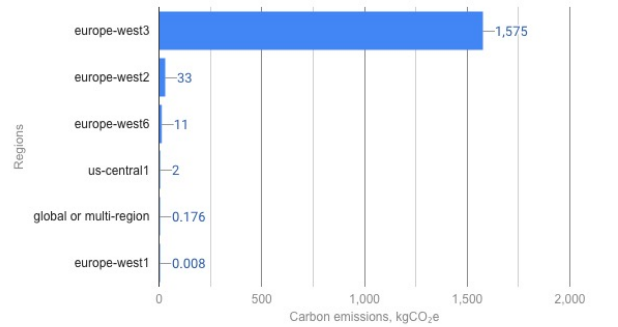
Gross carbon emissions by product in April 2022

Chart view



Gross carbon emissions by region in April 2022

Chart view



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 $kgCO_2eq$
Long-term storage	**$30 per TB-year**
	0.7 $kgCO_2eq$ per TB-year

Summary

The Arkivum Solution for ARCHIVER

NDSA Levels of Digital Preservation

- 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)

FAIRSFAIR
Fostering Fair Data Practices in Europe

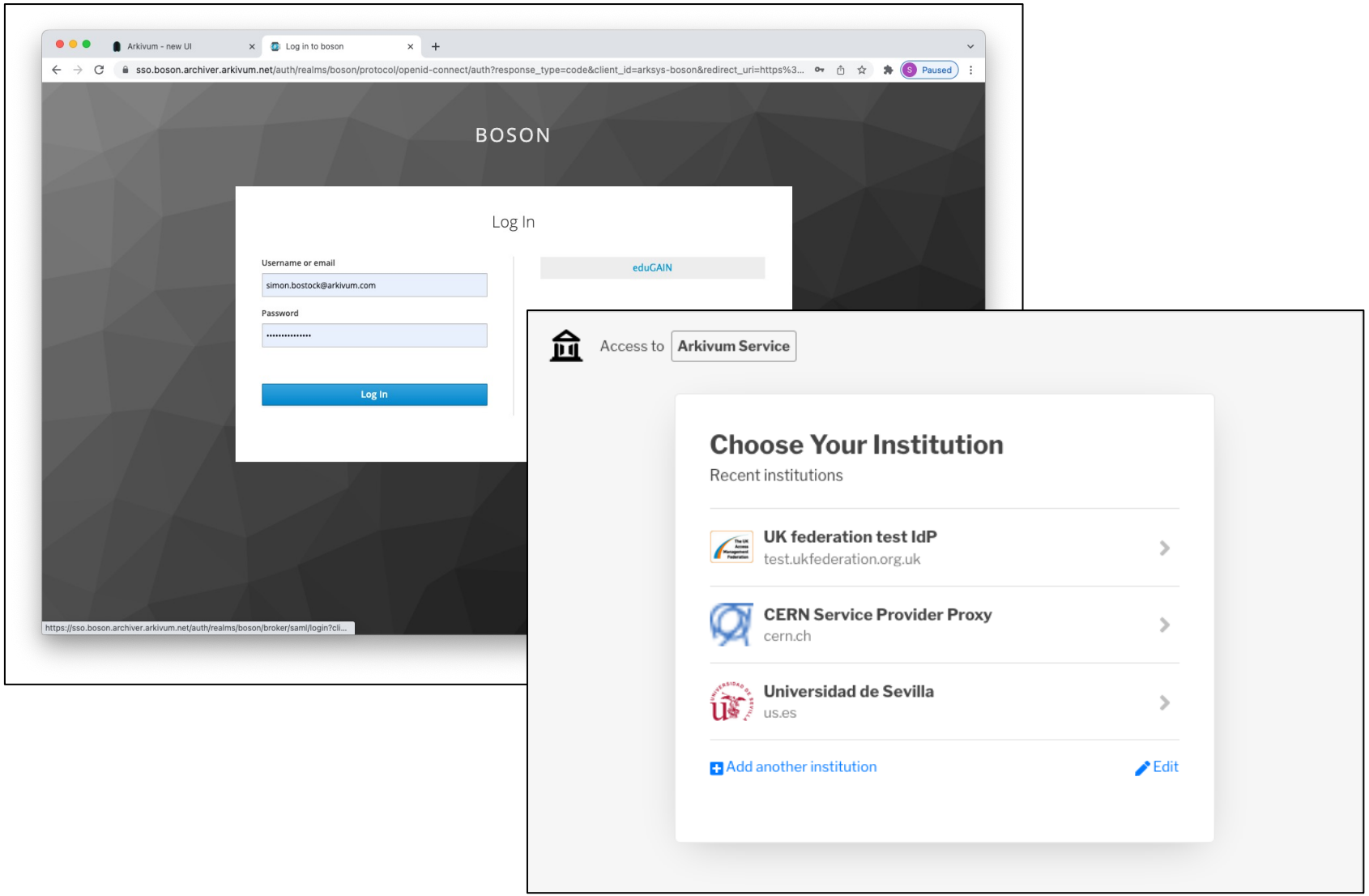
Solution: Part 1

Agenda

- Authentication and Authorisation (user interface and API)
- Overview of the Arkivum Web UI
- Configuration and Self-Service
- Getting data into the Arkivum solution

Authentication and Authorisation

- Local Accounts
- eduGAIN



Authorisation – User Roles

System Activities	Read Only	User	Superuser	Admin
Dashboard view of datapools	✅	✅	✅	✅
Search screen and any datasets associated to those assigned datapools	✅	✅	✅	✅
Download files and metadata		✅	✅	✅
Upload files into datapools (#1)		✅	✅	✅
Add and remove existing retentions to objects and collections		✅	✅	✅
Request preservation		✅	✅	✅
View retention rules (#1)		✅	✅	✅
Initiate bulk exports			✅	✅
Import and export metadata			✅	✅
Create retention rules				✅

(#1) – Access to the respective datapool(s) assigned

Authorisation – User Roles

Reporting	Read Only	User	Superuser	Admin
Ingest Report		✅	✅	✅
Preservation Report		✅	✅	✅
Normalisation Report		✅	✅	✅
Hold Report		✅	✅	✅
Unhold Report		✅	✅	✅
Processing Report			✅	✅
Export Report			✅	✅
Audit Trail			✅	✅
Deletion Report				✅

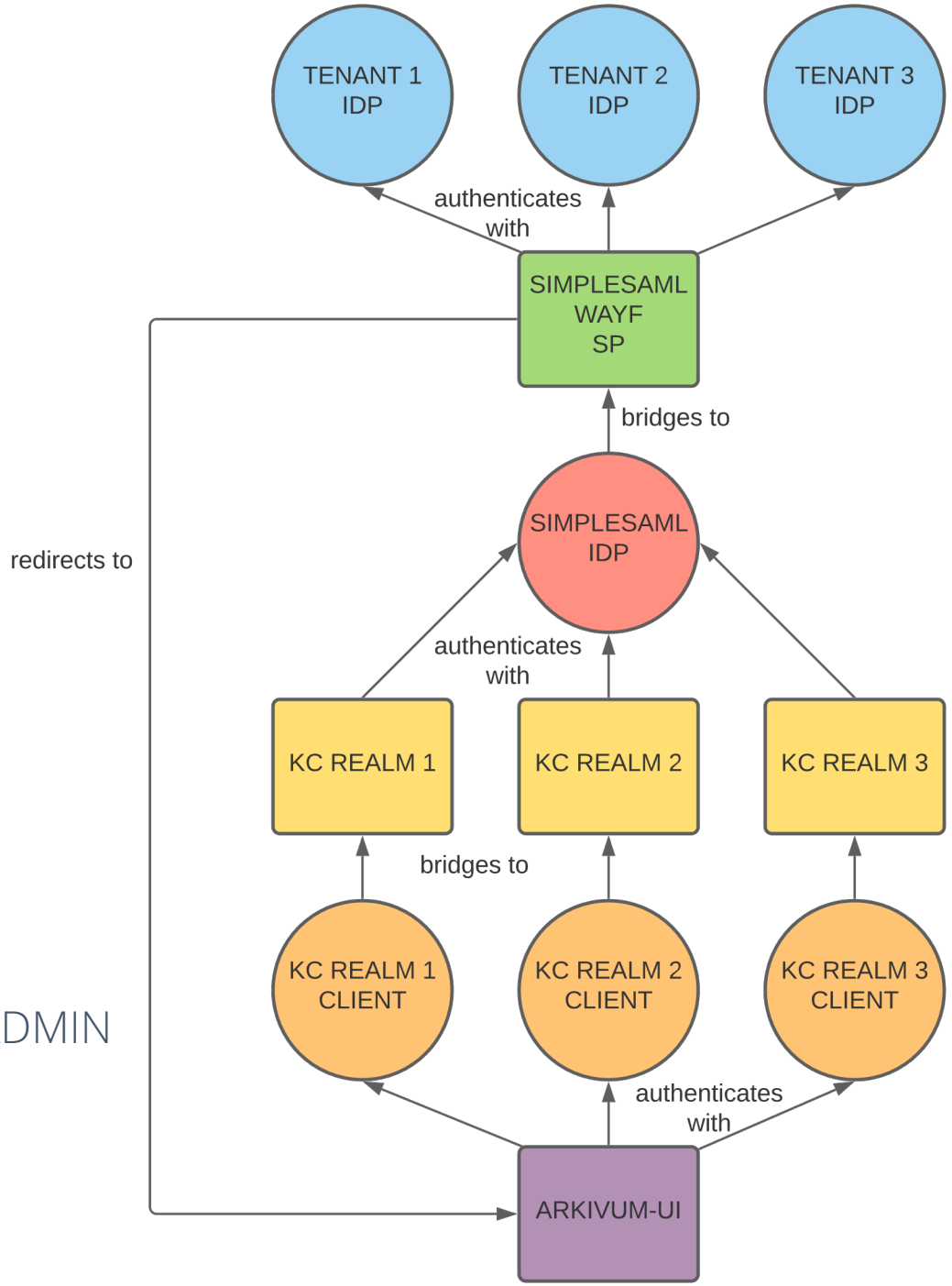
Approval Workflows	Read Only	User	Superuser	Admin
Approve/reject ingests		✅	✅	✅
Approve/reject deletion requests		✅	✅	✅
Approve/reject Bulk Exports			✅	✅

Authentication and Authorisation

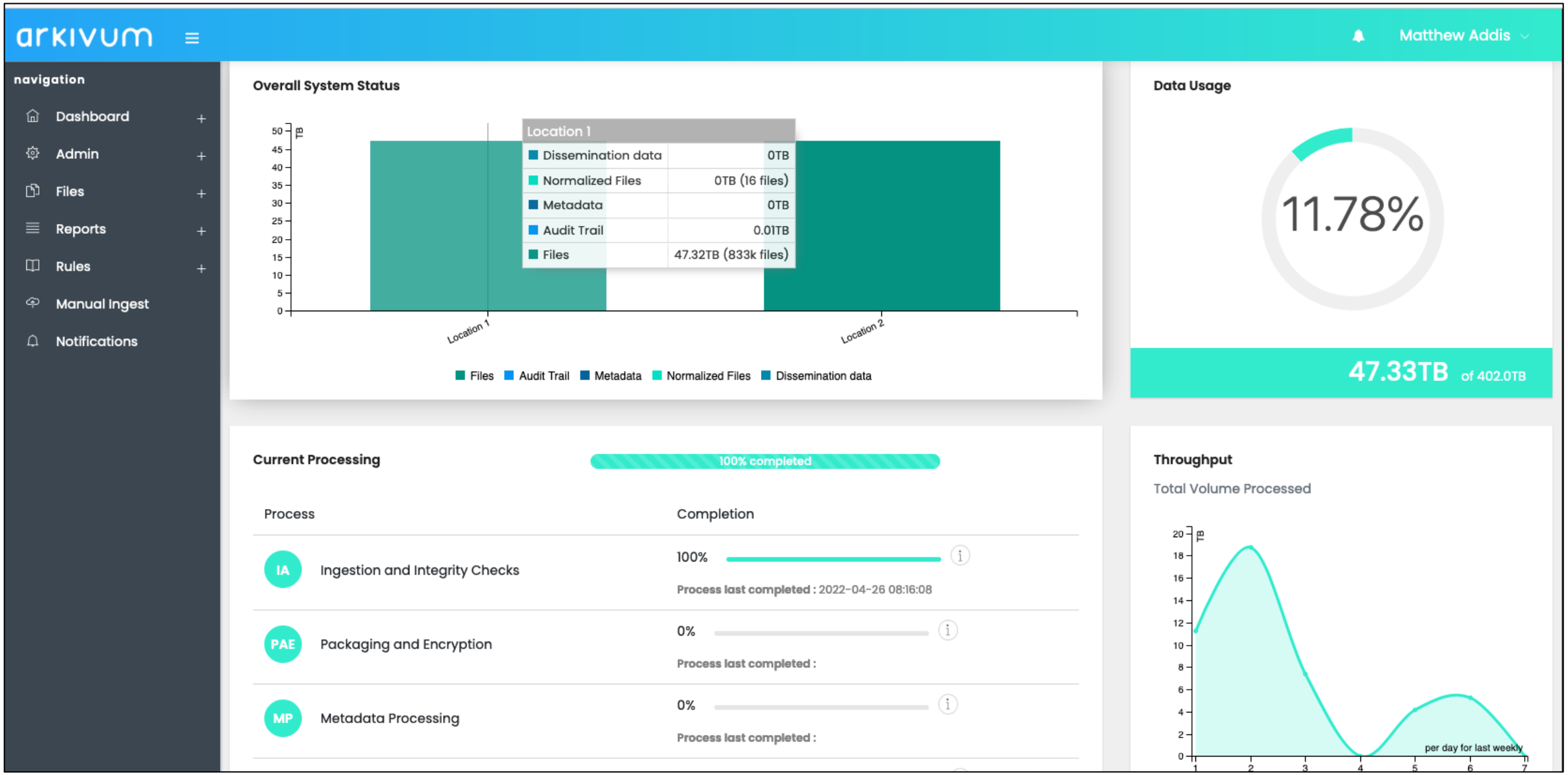
https://vimeo.com/723625238/b5a3ee264a

AAI Summary

- Arkivum is a member of the UKAMF
- Arkivum is a registered SP
  - REFEDS R&S, DP CoCo, Sirtfi
- Implementation using simpleSAML and keycloak
- WAYF allows users to select their home IdP
- Routing of authenticated users to specific tenants
  - e.g. user from CERN IdP -> cern.archiver.arkivum.net
- Mapping of user attributes to roles
  - e.g. User X with attribute Y that has value Z -> ROLE_ADMIN
  - eduPersonEntitlement (part of REFEDS R&S bundle)



Dashboard Overview



Dashboard Overview

https://vimeo.com/722917823/17c05d473a

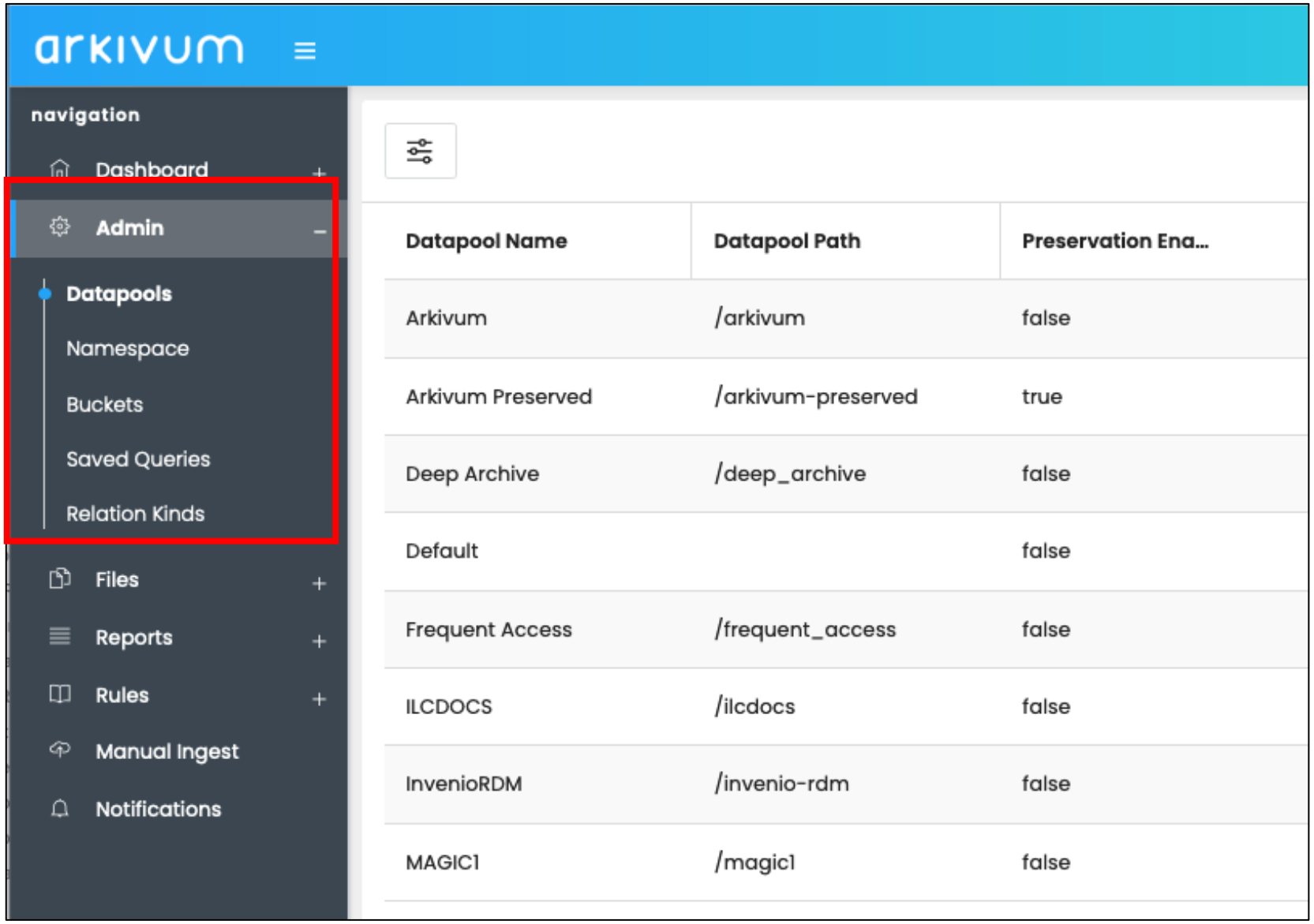
Arkivum Dashboard

- System usage: data volumes, throughput, processes
- Administration: datapools, metadata configuration, buckets, retentions and holds
- Ingest and Export: via buckets, direct upload/download
- Search and Navigation: browse and find datasets, files, records
- Reports: ingest, preservation, exports, deletion, retentions, holds, audit log
- Notifications: approvals, alerts

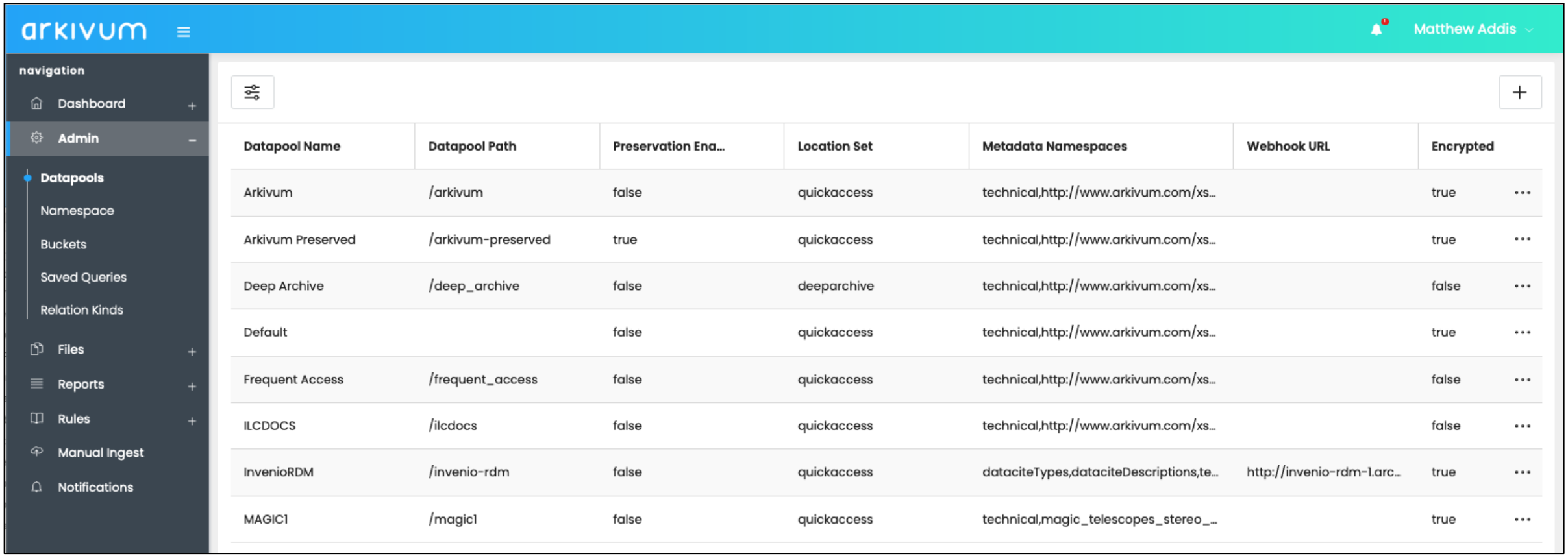
- Web based, cross-platform
- Everything in the UI is also available through a REST API

Configuration

- Datapools
- Metadata namespaces
- Buckets



Datapools

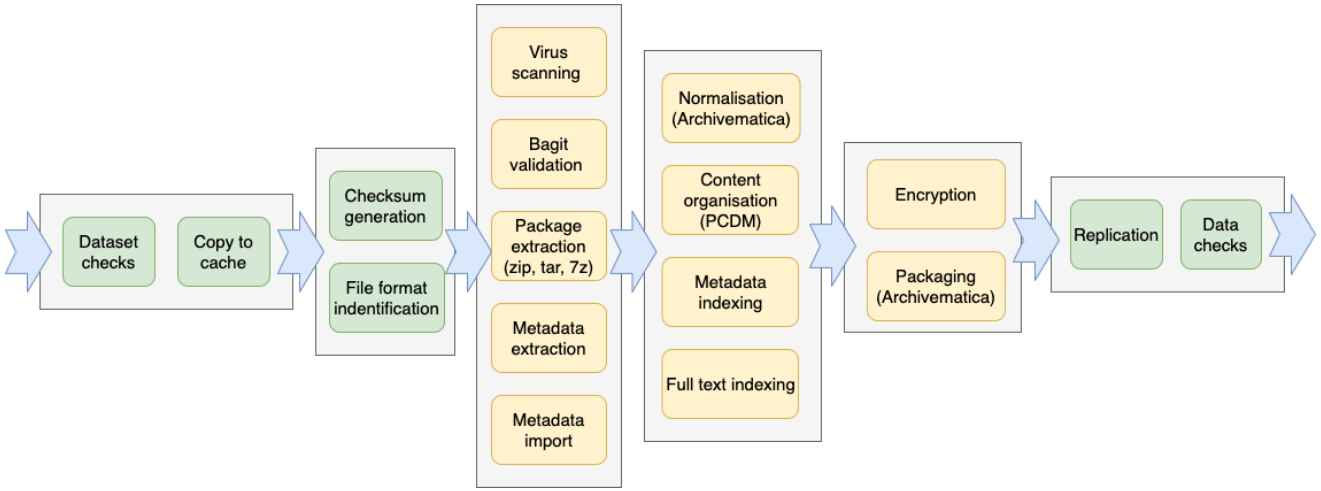


Datapool Name	Datapool Path	Preservation Ena...	Location Set	Metadata Namespaces	Webhook URL	Encrypted
Arkivum	/arkivum	false	quickaccess	technical,http://www.arkivum.com/xs...		true
Arkivum Preserved	/arkivum-preserved	true	quickaccess	technical,http://www.arkivum.com/xs...		true
Deep Archive	/deep_archive	false	deeparchive	technical,http://www.arkivum.com/xs...		false
Default		false	quickaccess	technical,http://www.arkivum.com/xs...		true
Frequent Access	/frequent_access	false	quickaccess	technical,http://www.arkivum.com/xs...		false
ILCDOCS	/ilcdocs	false	quickaccess	technical,http://www.arkivum.com/xs...		false
InvenioRDM	/invenio-rdm	false	quickaccess	dataciteTypes,dataciteDescriptions,te...	http://invenio-rdm-1.arc...	true
MAGIC1	/magic1	false	quickaccess	technical,magic_telescopes_stereo_...		true

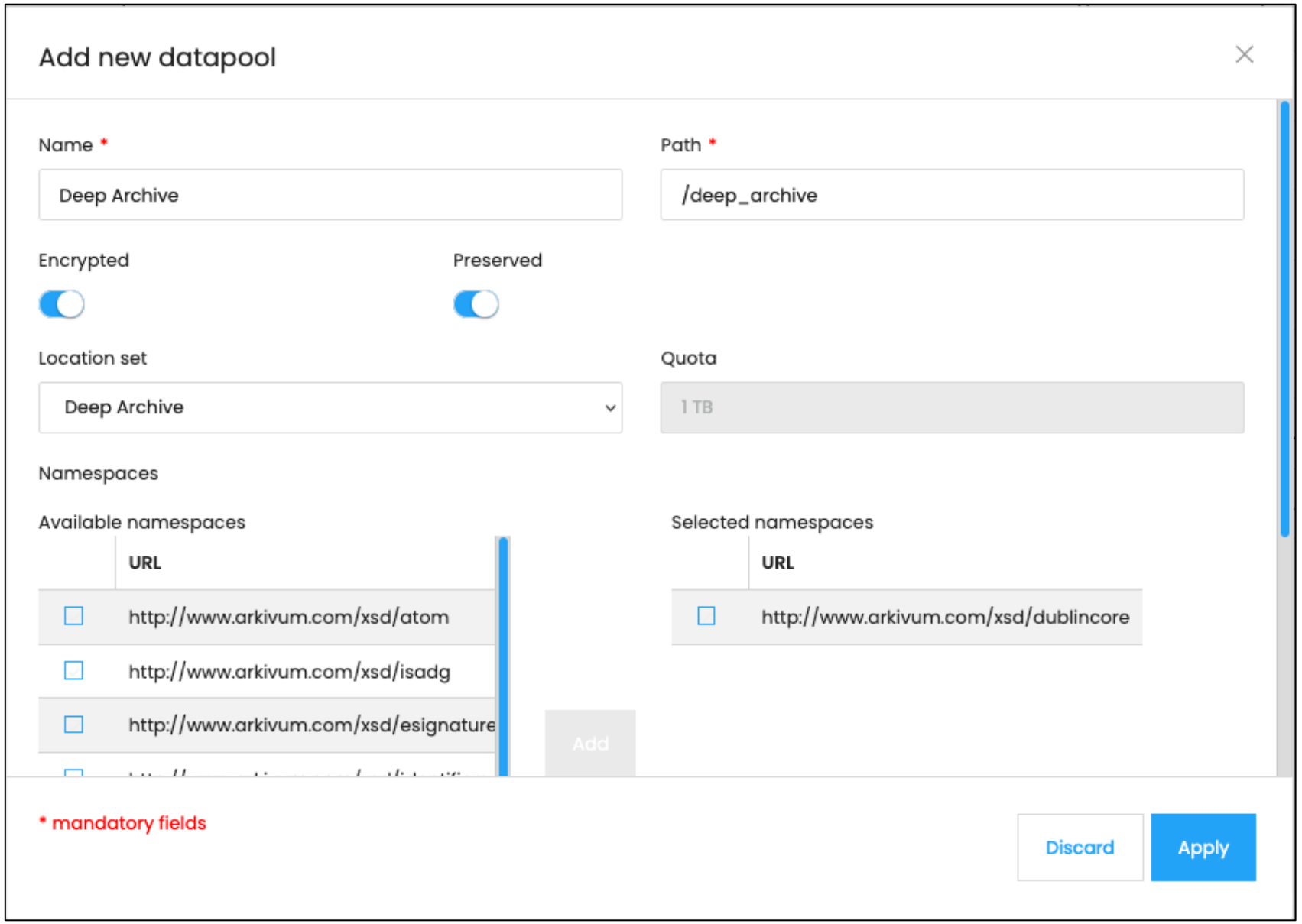
Datapool	Safeguarding	Preservation	InvenioRDM	Escrow	Read Only Access	User Access	Admin Access
Datapool 01	✅	✅		✅	User 1	User Group 1	Admin Group
Datapool 02	✅	✅	✅	✅	User 2	User Group 2	Admin Group
Datapool 03	✅		✅	✅		User Group 3	Admin Group

Datapools

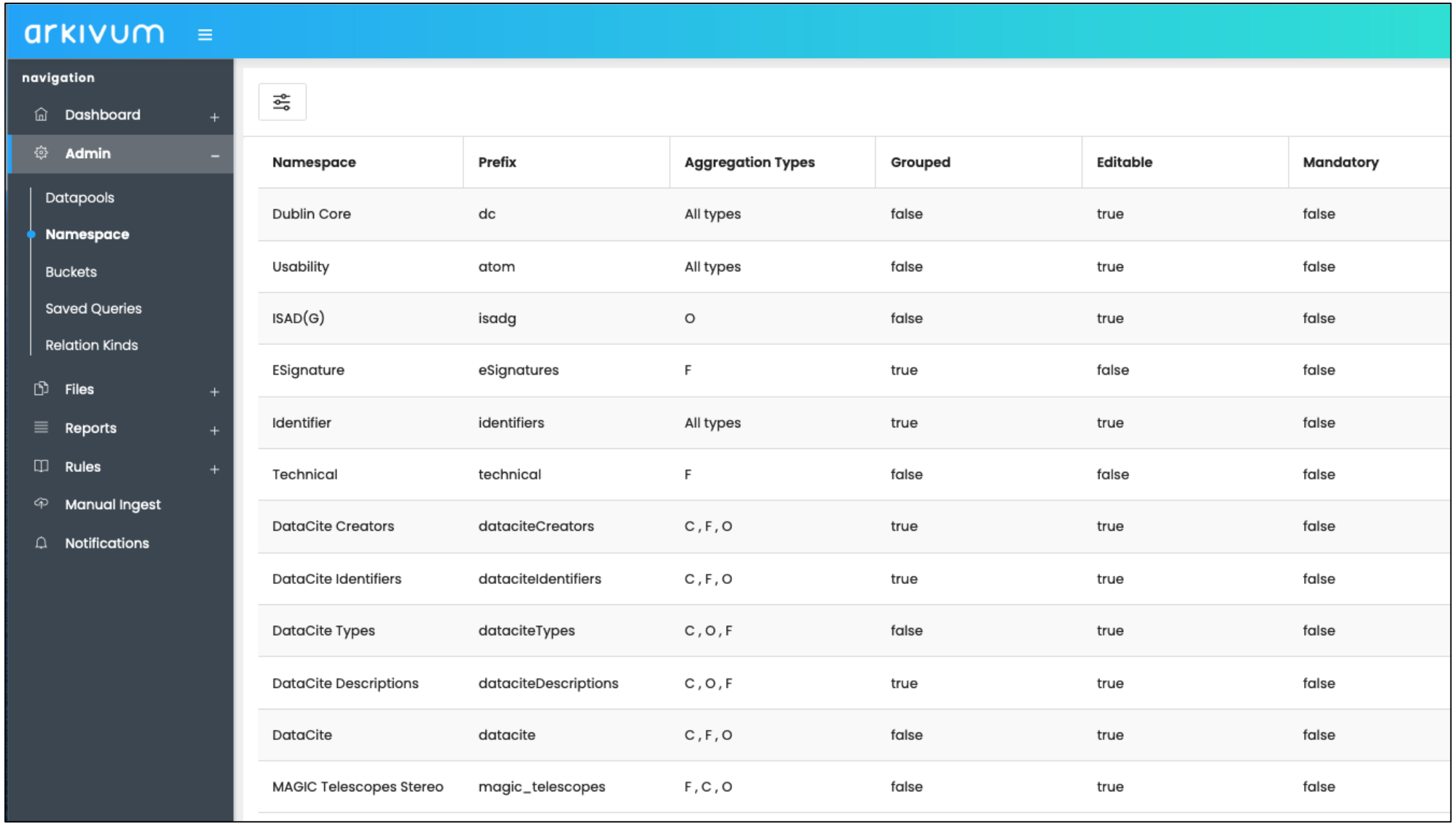
- Where data will be stored for the long-term
  - Frequent access or deep archive buckets
- What metadata fields and rules should be applied
  - Dublin Core, DataCite, domain specific
- What processes to run when data is ingested
  - Safeguarding and preservation
- Automatic cleanup
  - E.g. remove datasets after successful ingest
- File encryption
  - In addition to cloud storage, user provided keys
- User access
  - Users can be given roles that allow access to specific datasets
- Webhooks
  - Called at the end of the ingest and/or preservation process



Datapools


Add new datapool
Name *
Deep Archive
Path *
/deep_archive
Encrypted
Preserved
Location set
Deep Archive
Quota
1 TB
Namespaces
Available namespaces
URL
http://www.arkivum.com/xsd/atom
http://www.arkivum.com/xsd/isadg
http://www.arkivum.com/xsd/esignature
Add
Selected namespaces
URL
http://www.arkivum.com/xsd/dublincore
* mandatory fields
Discard
Apply

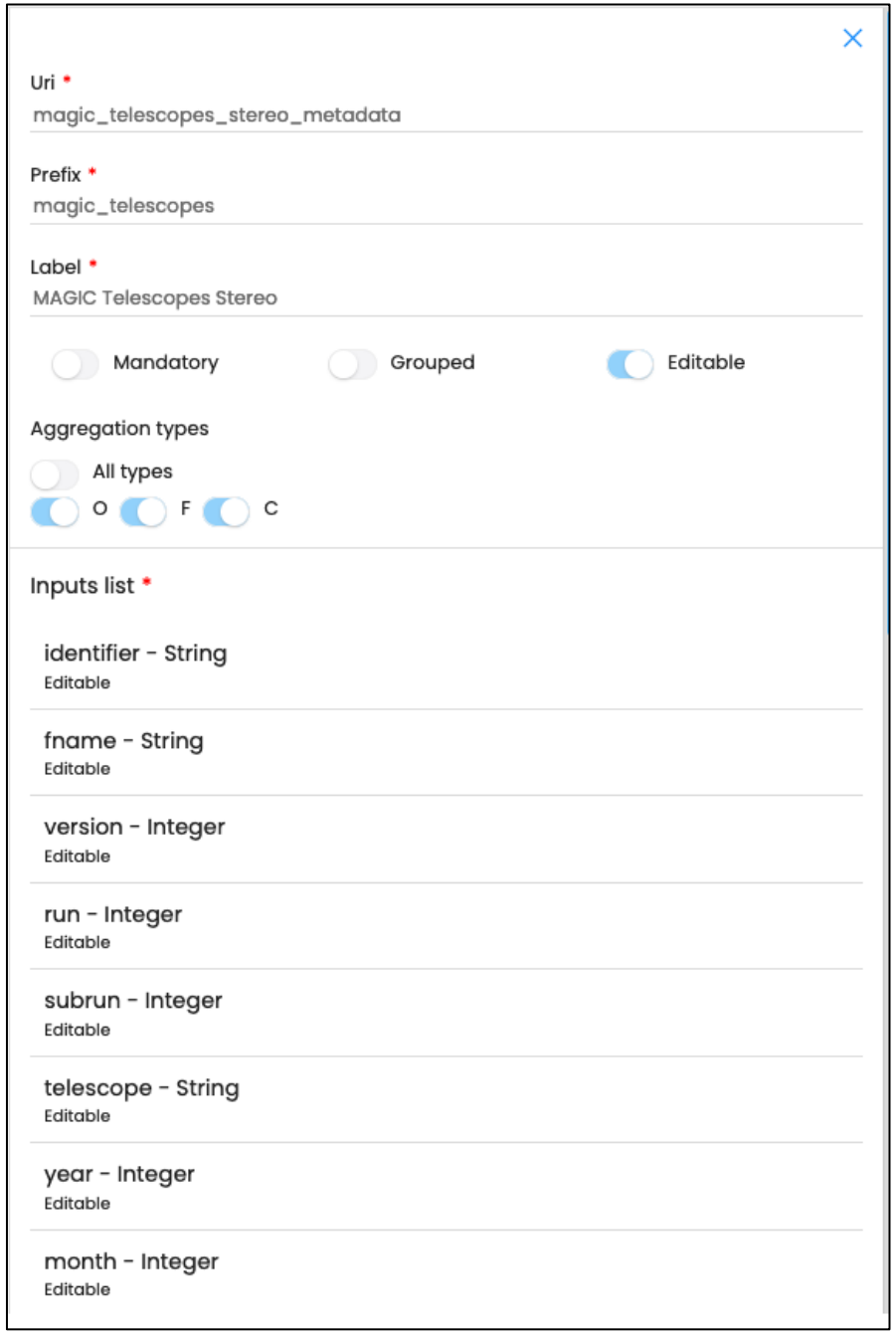
Metadata Namespaces



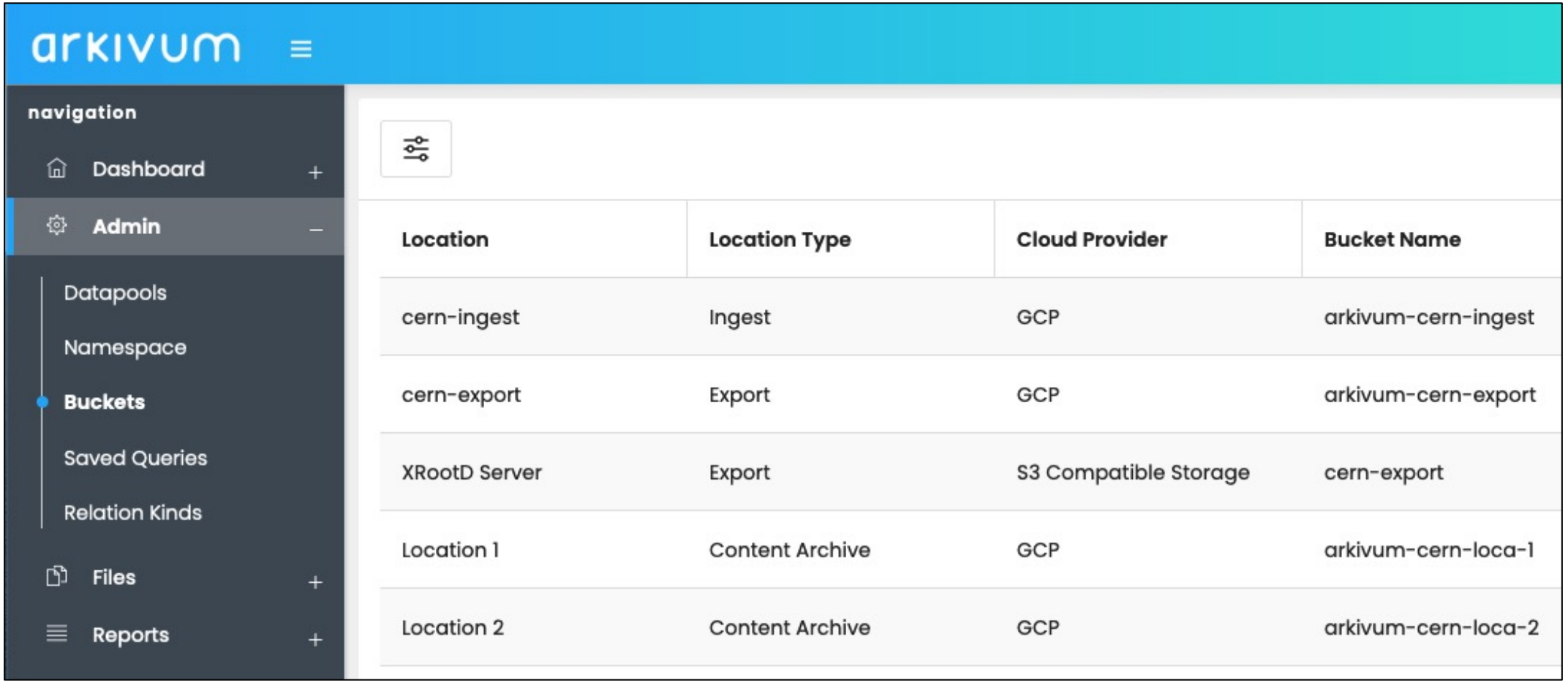
Namespace	Prefix	Aggregation Types	Grouped	Editable	Mandatory
Dublin Core	dc	All types	false	true	false
Usability	atom	All types	false	true	false
ISAD(G)	isadg	O	false	true	false
ESignature	eSignatures	F	true	false	false
Identifier	identifiers	All types	true	true	false
Technical	technical	F	false	false	false
DataCite Creators	dataciteCreators	C , F , O	true	true	false
DataCite Identifiers	dataciteIdentifiers	C , F , O	true	true	false
DataCite Types	dataciteTypes	C , O , F	false	true	false
DataCite Descriptions	dataciteDescriptions	C , O , F	true	true	false
DataCite	datacite	C , F , O	false	true	false
MAGIC Telescopes Stereo	magic_telescopes	F , C , O	false	true	false

Metadata Namespaces

- Fields
  - Title, Subject, Date etc.
- Field Types
  - String, Date, Integer
- Repeatable or Single
  - Titles, Identifiers etc.
- Grouped
  - First Name, Last Name, ORCID, Affiliation
- Editable
  - Write once or updateable by users



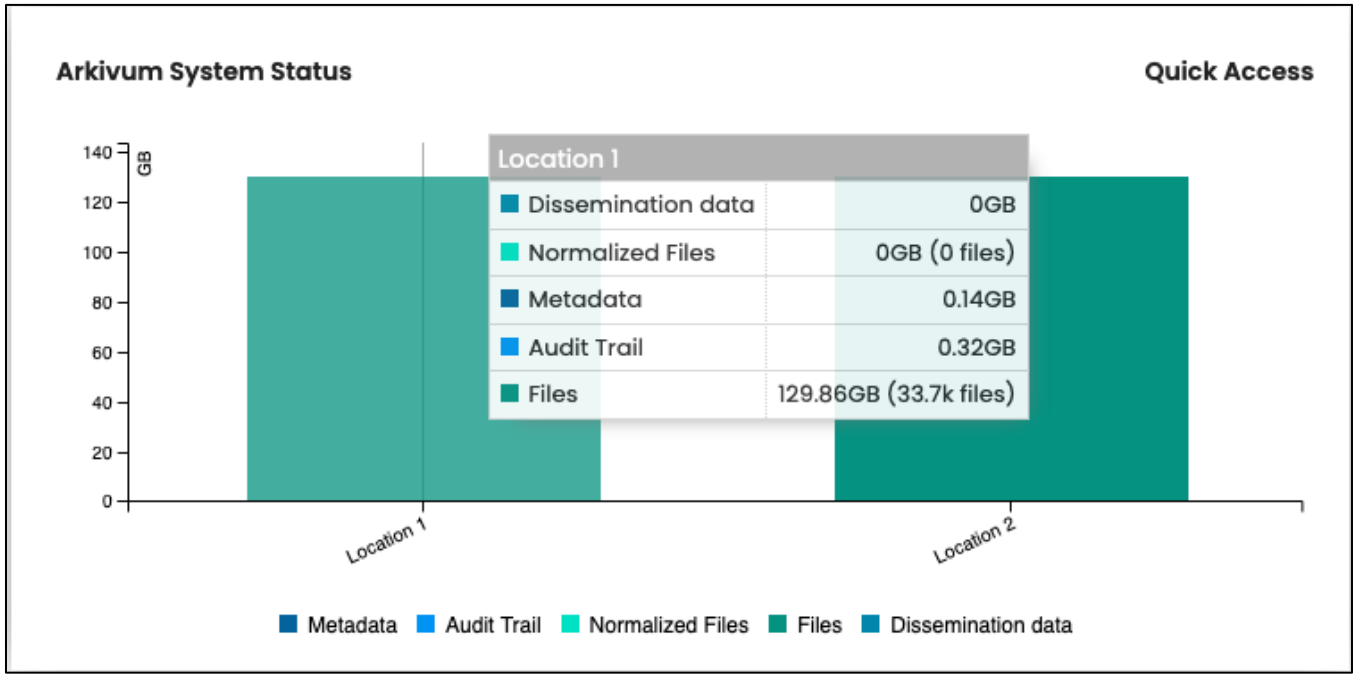
Buckets

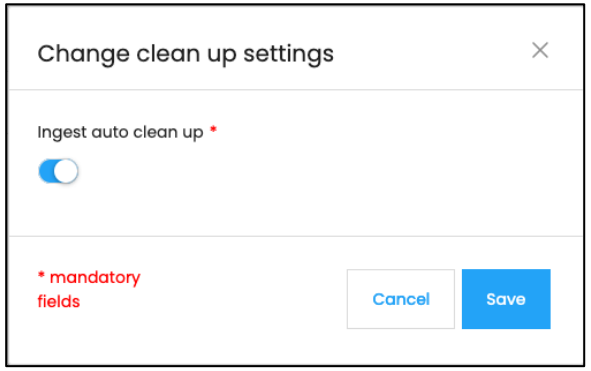


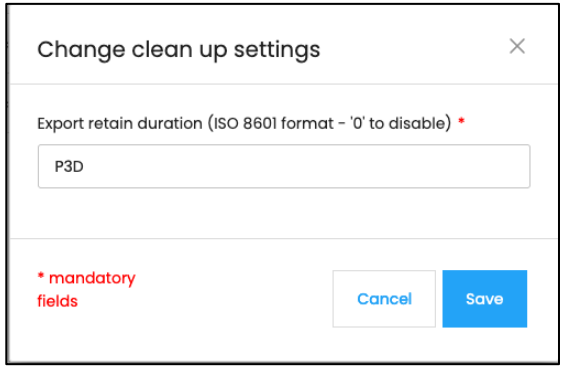
Location	Location Type	Cloud Provider	Bucket Name
cern-ingest	Ingest	GCP	arkivum-cern-ingest
cern-export	Export	GCP	arkivum-cern-export
XRootD Server	Export	S3 Compatible Storage	cern-export
Location 1	Content Archive	GCP	arkivum-cern-loca-1
Location 2	Content Archive	GCP	arkivum-cern-loca-2

Buckets

- Ingest, Storage, Export
- AWS, GCP, on-prem
  - GCP: standard, nearline, coldline, archive
  - AWS: standard, glacier, deep glacier
- Provided by Arkivum or external
- Automatic cleanup of files
- Grouped together for long-term storage
  - E.g. 1 copy GCP frequent access, 1 copy GCP deep archive, 1 copy Azure escrow



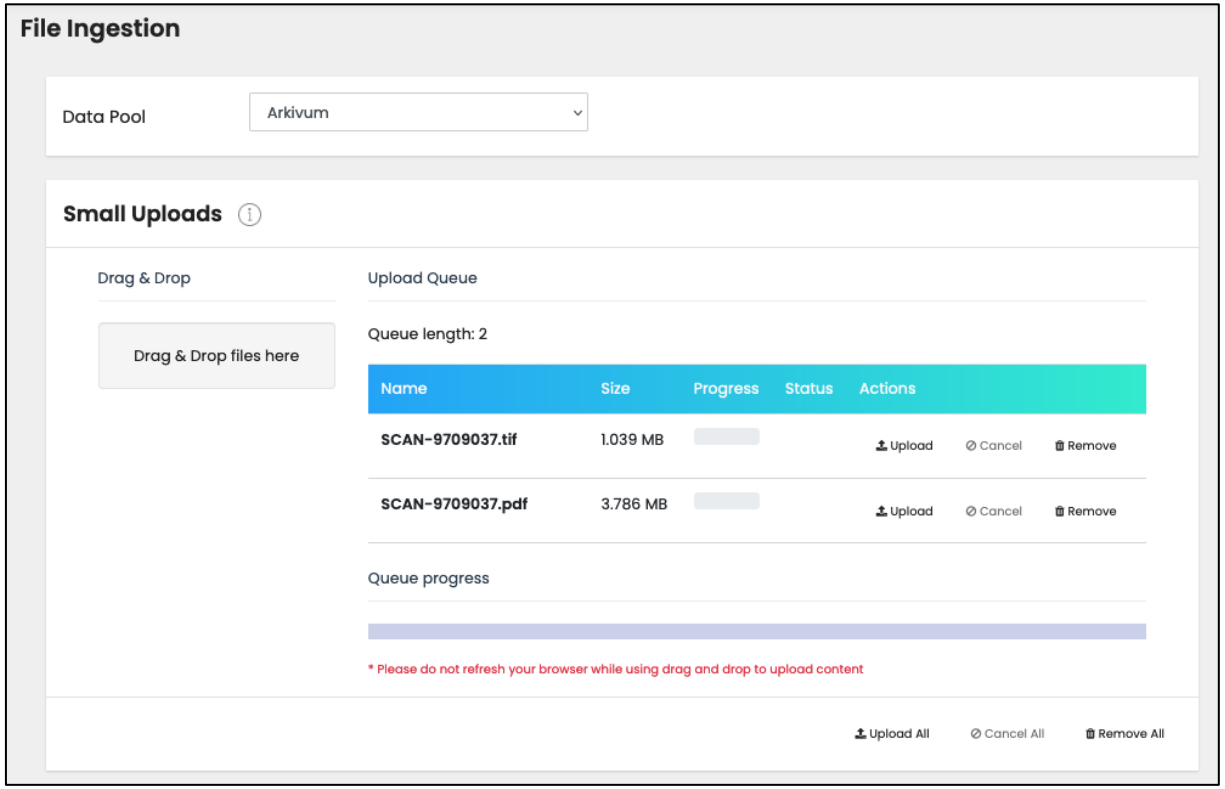




Ingest of Datasets

- Ingest via buckets
- Ingest via REST API
- Ingest via Web UI

- Optional use of bagit
- Optional use of archive containers (tar, zip, 7z)
- Optional ingest of metadata



RCLONE

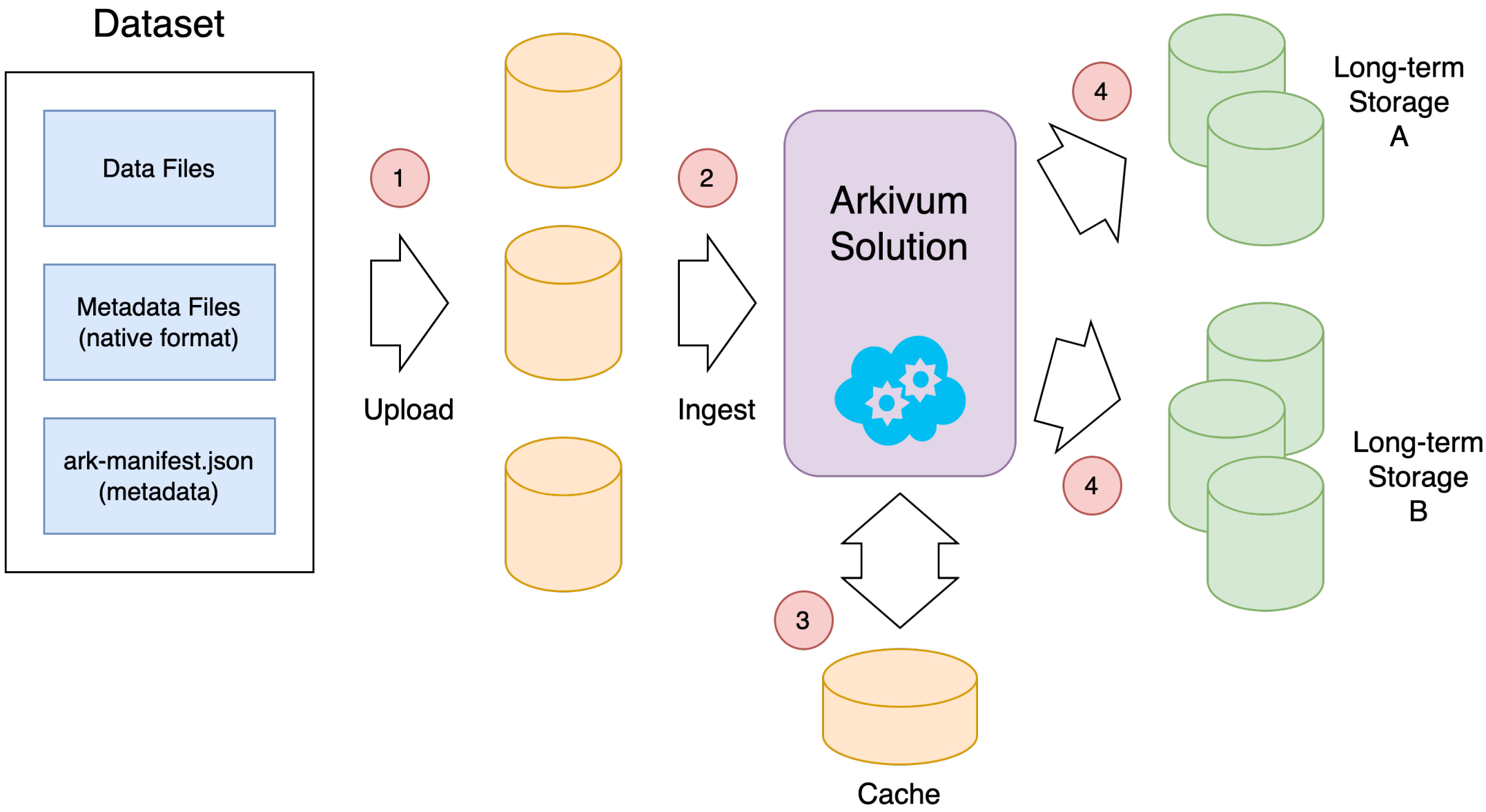
GCP

Amazon S3

research object.org

Bagit & BDbags

Ingest Workflow using Buckets

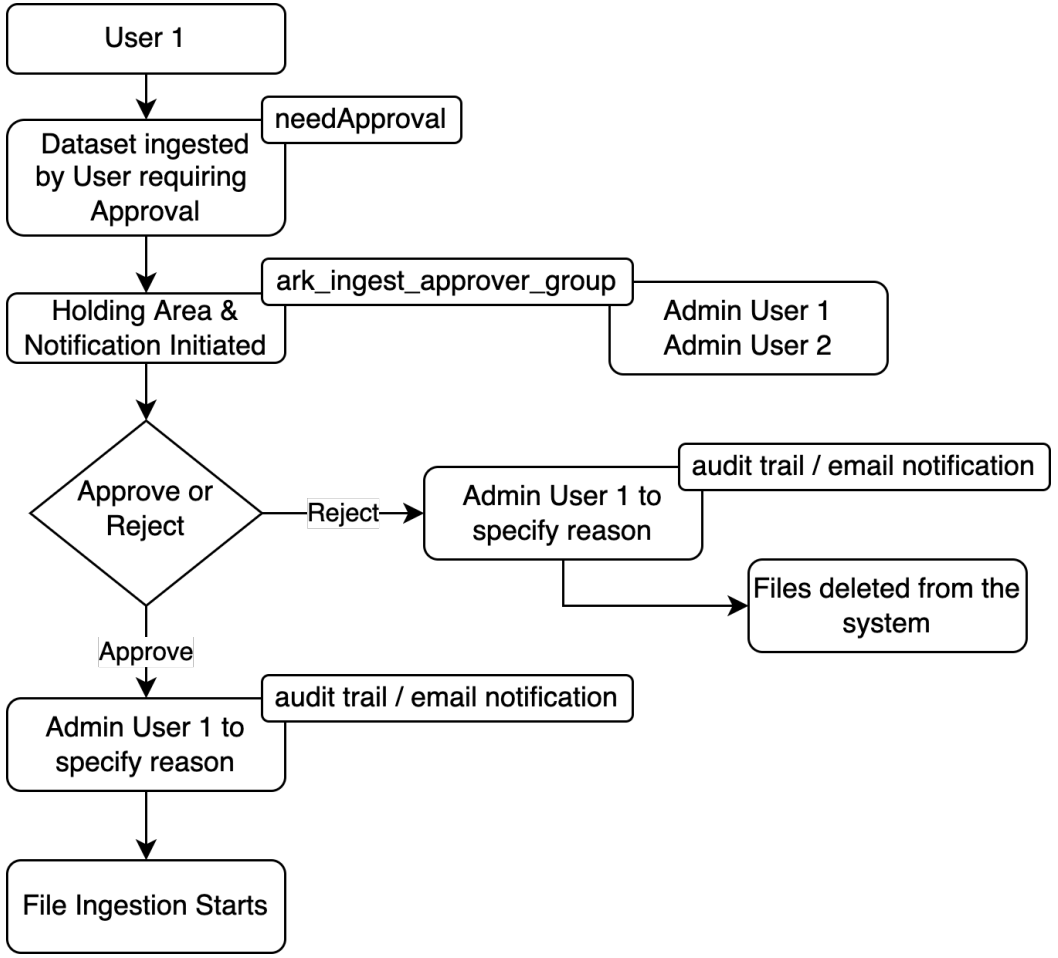


Dataset Ingest

https://vimeo.com/722918646/46a9059c74

Optional Ingest Approval Workflow

- Users can upload content, but not ingest it
- Ingest Approvers get notified of new datasets
- Approvers can review datasets and approve/reject
- Datasets ingested
- Decisions and actions in audit trail
- Notifications when ingest complete

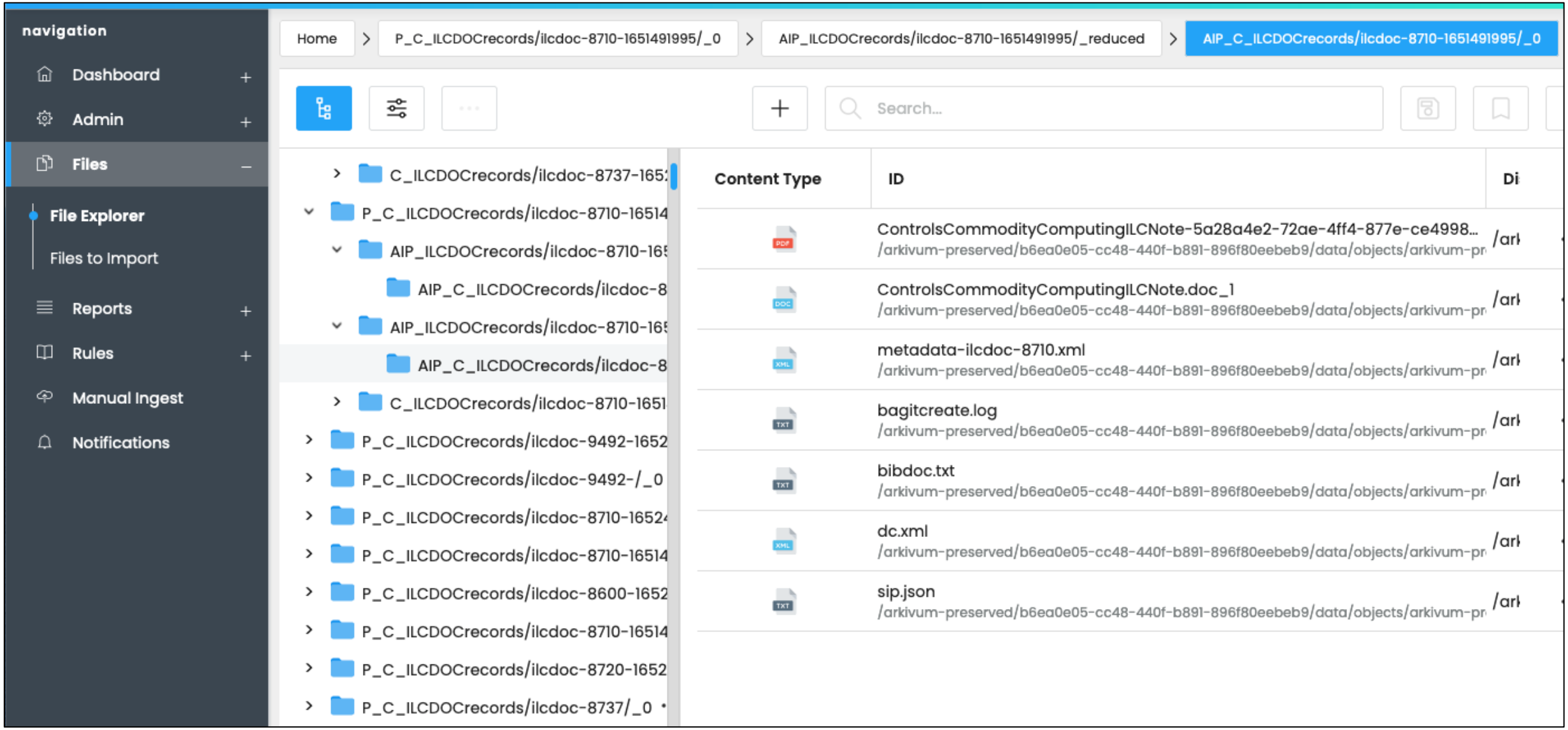


Solution: Part 2

Agenda

- Search and Navigation
- Export of data and metadata
- Providing and updating metadata

Search and Navigation

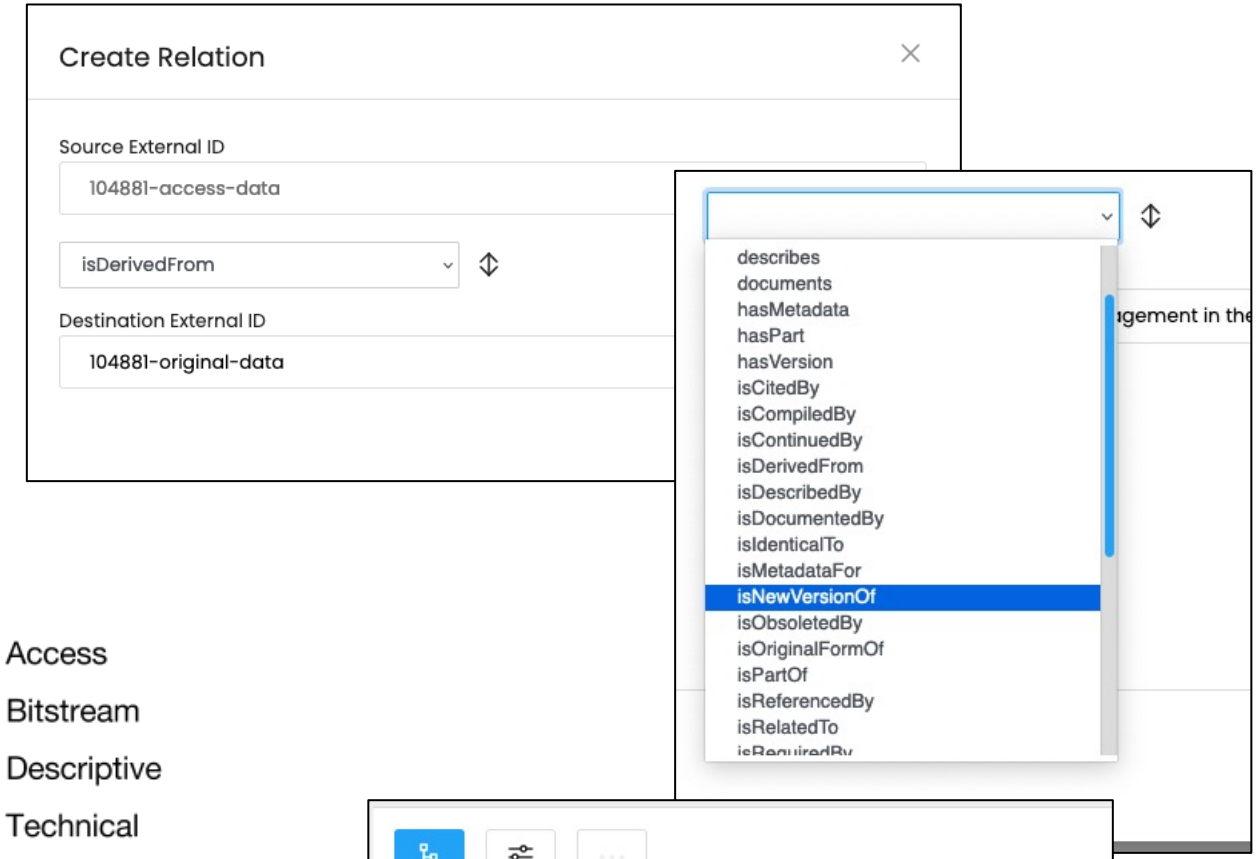


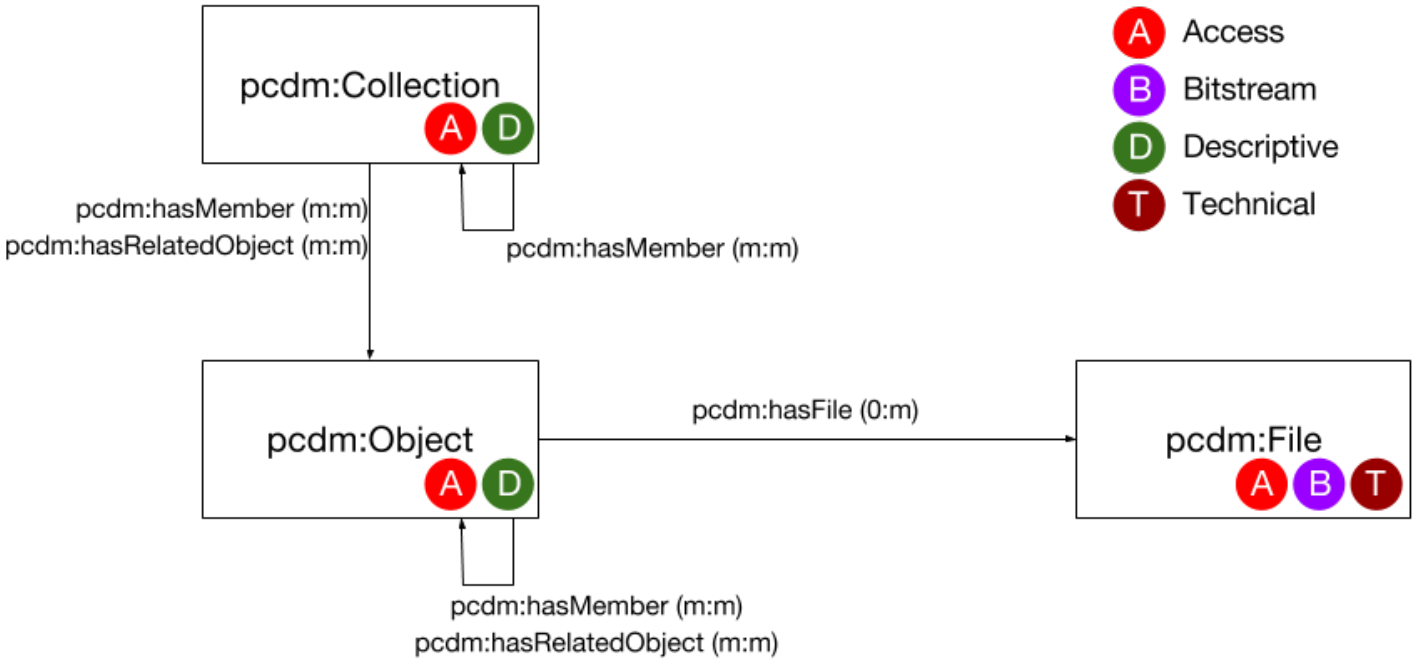
Search and Navigation

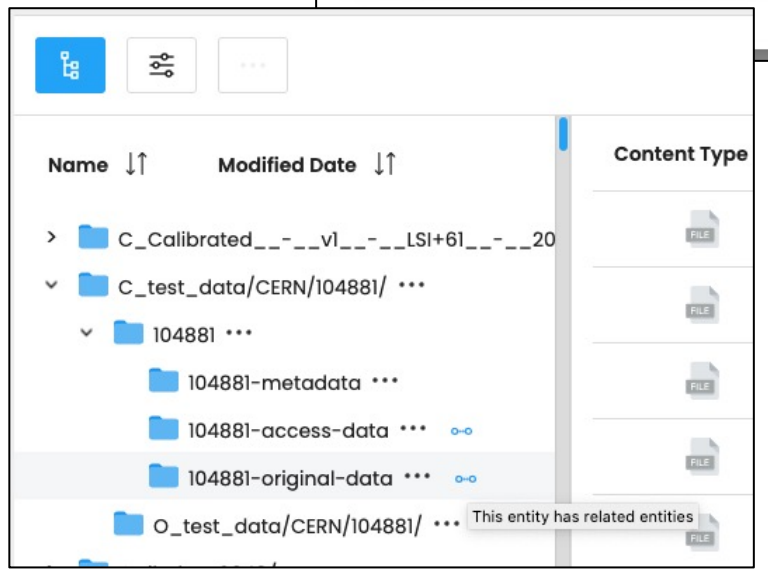
https://vimeo.com/723627686/2153b90960

Hierarchical Data Structures

- Datasets can be structured using PCDM
- Structure defined in metadata
- Tree viewer shows hierarchies and links
- Breadcrumbs and other navigation aids

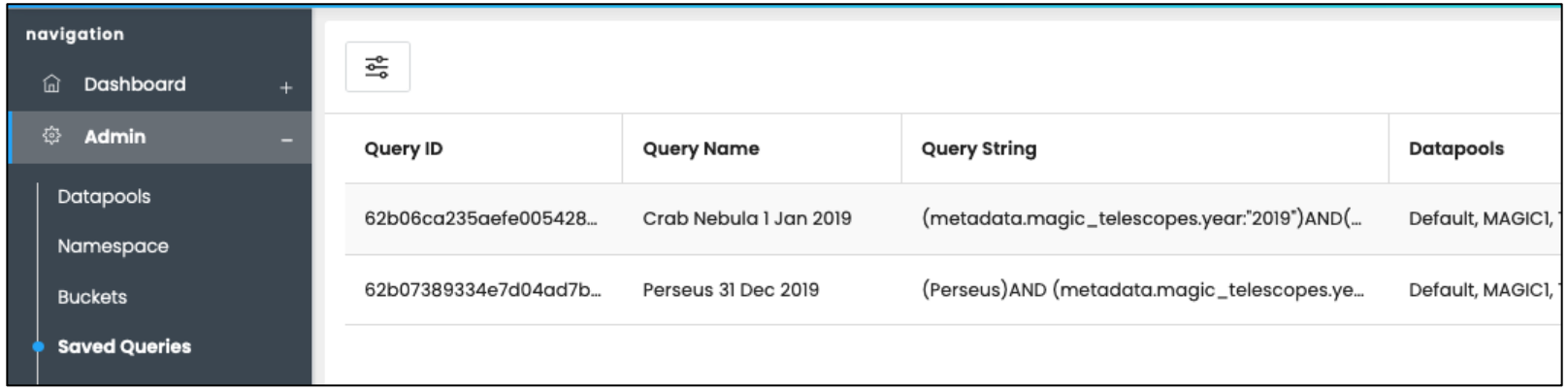






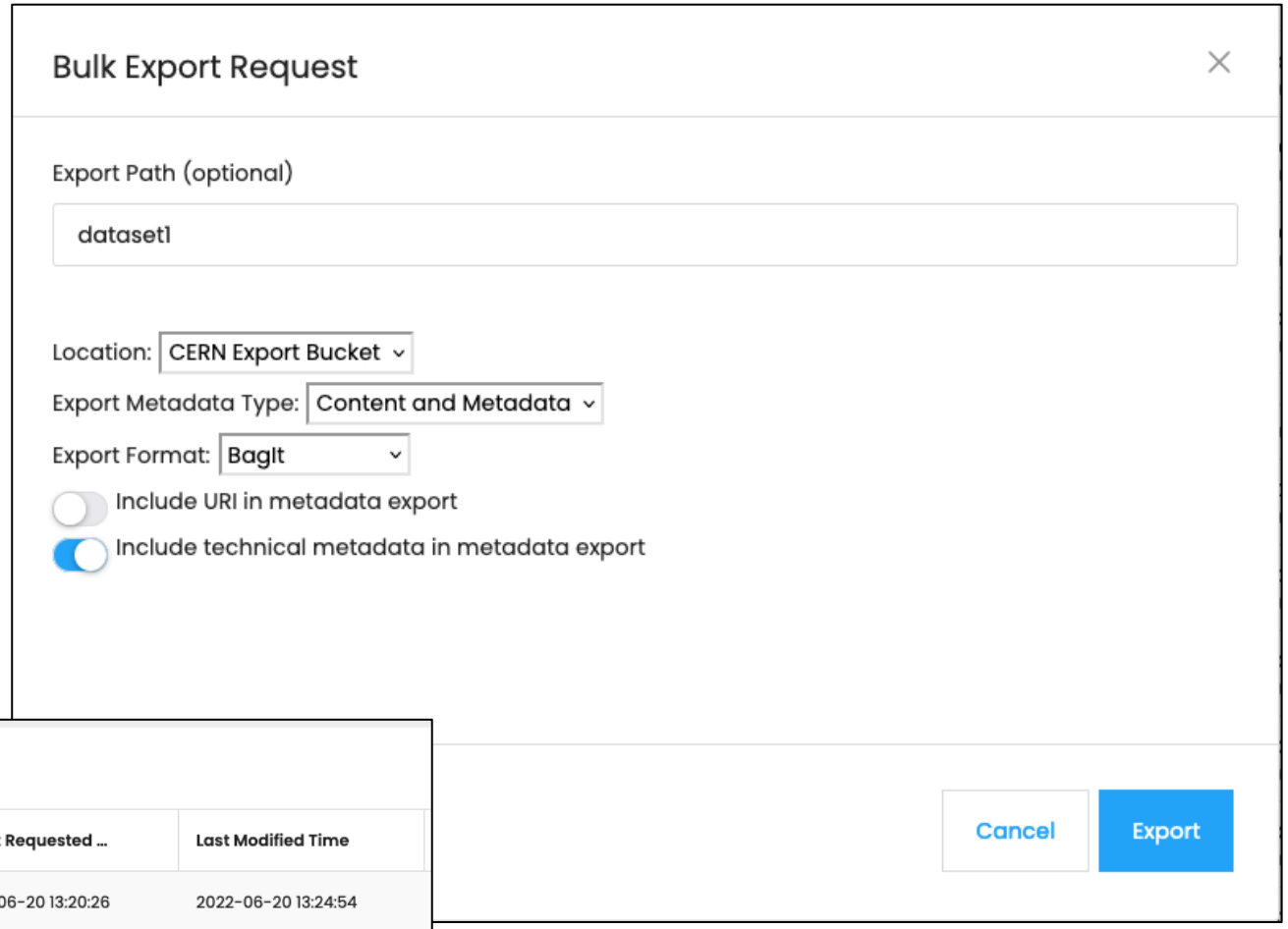
Searching

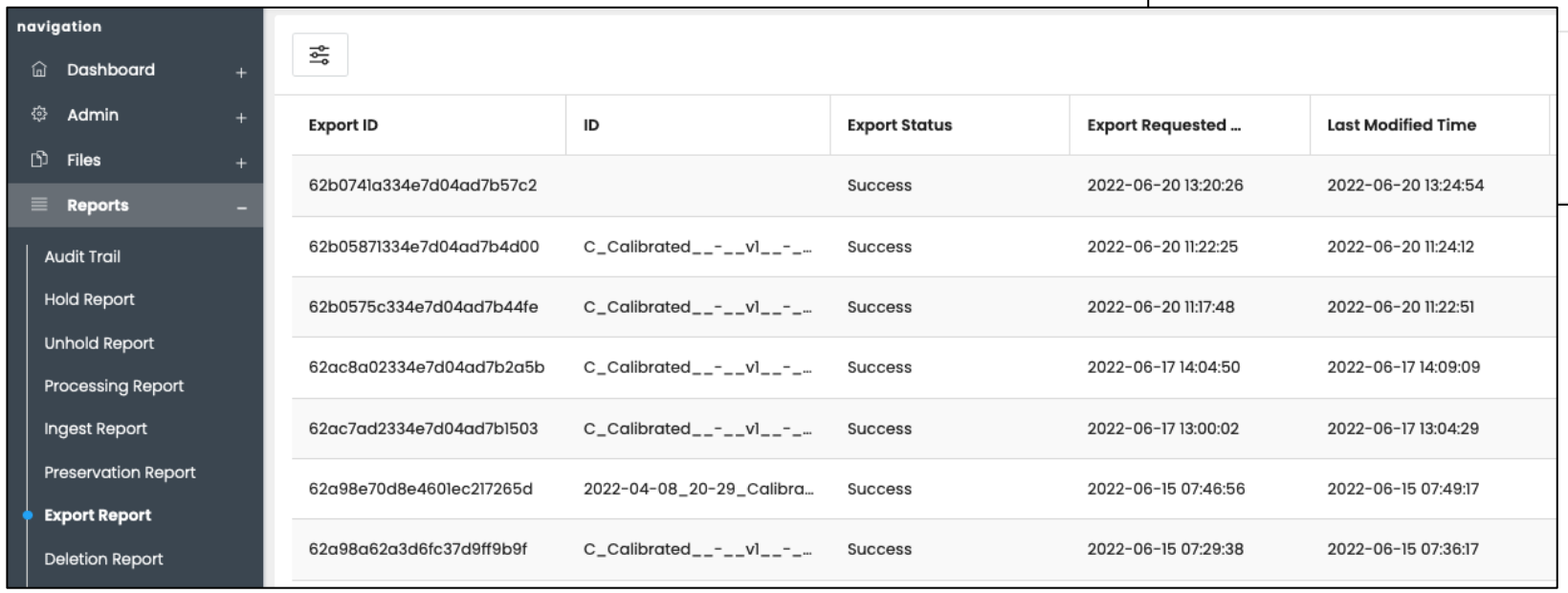
- Search box and search expressions (Booleans, fuzzy matches, wildcards etc.)
- Query builder
- Filter by datapools, fields, type of entity (files, collections etc.)
- Searches only return results that user is entitled to see
- Save and re-run queries
- Easy export results of a search
- REST API or UI
- Backed by ElasticSearch allowing use of DSL or Query Strings



Export

- Files and/or Metadata
- With/without bagit
- Choice of buckets
- Webhooks and Notifications
- Request/Approve workflow
- Export reports and audit trail





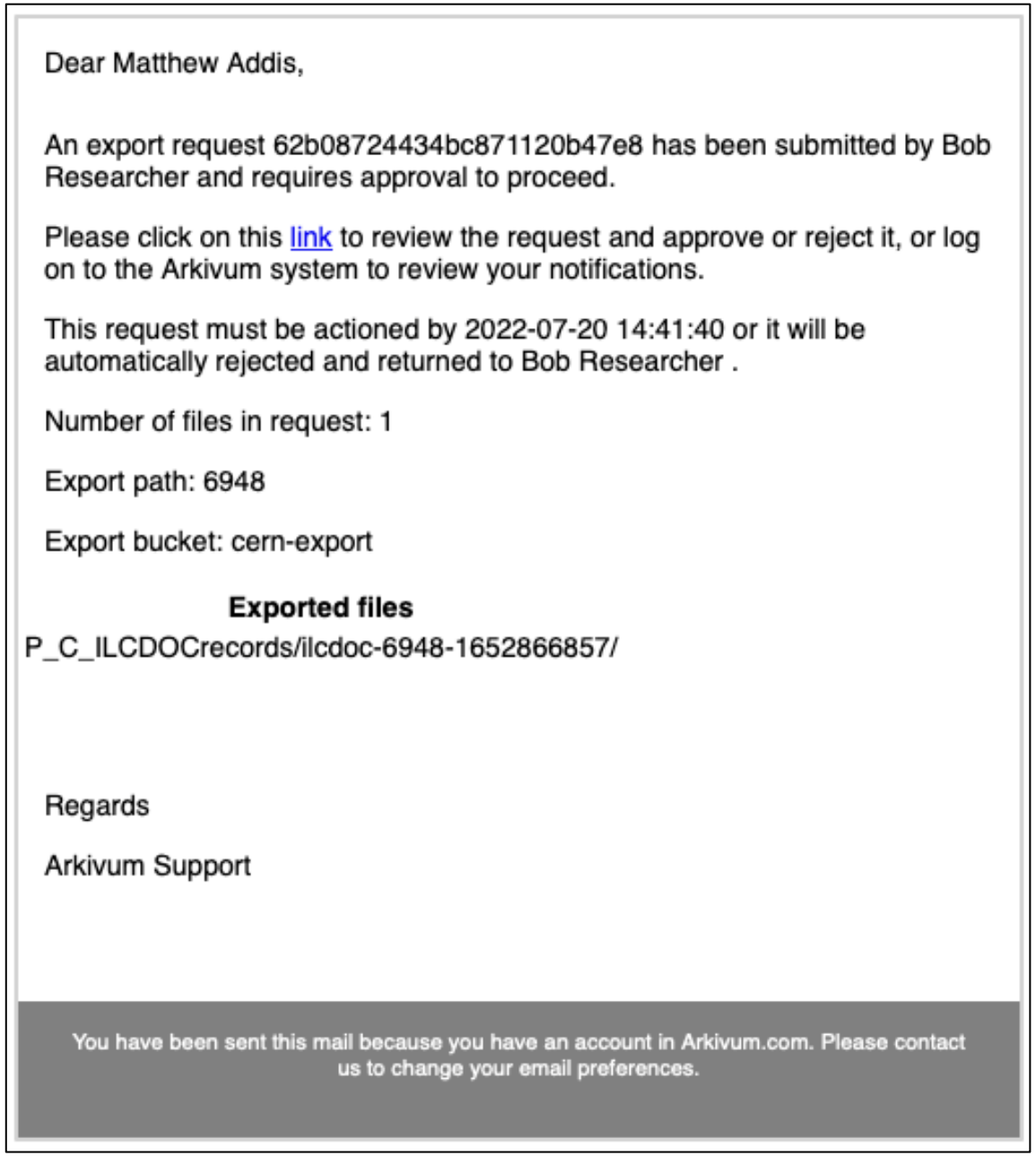
Export ID	ID	Export Status	Export Requested ...	Last Modified Time
62b0741a334e7d04ad7b57c2		Success	2022-06-20 13:20:26	2022-06-20 13:24:54
62b05871334e7d04ad7b4d00	C_Calibrated__-__v1__-_...	Success	2022-06-20 11:22:25	2022-06-20 11:24:12
62b0575c334e7d04ad7b44fe	C_Calibrated__-__v1__-_...	Success	2022-06-20 11:17:48	2022-06-20 11:22:51
62ac8a02334e7d04ad7b2a5b	C_Calibrated__-__v1__-_...	Success	2022-06-17 14:04:50	2022-06-17 14:09:09
62ac7ad2334e7d04ad7b1503	C_Calibrated__-__v1__-_...	Success	2022-06-17 13:00:02	2022-06-17 13:04:29
62a98e70d8e4601ec217265d	2022-04-08_20-29_Calibra...	Success	2022-06-15 07:46:56	2022-06-15 07:49:17
62a98a62a3d6fc37d9ff9b9f	C_Calibrated__-__v1__-_...	Success	2022-06-15 07:29:38	2022-06-15 07:36:17

Export

https://vimeo.com/723623926/b82a560f3d

Export Approval Workflow

- Optional workflow
- Allows control over costs and access
- Can be combined with integrations, e.g. xrootd and InvenioRDM



Dear Matthew Addis,

An export request 62b08724434bc871120b47e8 has been submitted by Bob Researcher and requires approval to proceed.

Please click on this link to review the request and approve or reject it, or log on to the Arkivum system to review your notifications.

This request must be actioned by 2022-07-20 14:41:40 or it will be automatically rejected and returned to Bob Researcher .

Number of files in request: 1

Export path: 6948

Export bucket: cern-export

Exported files

P_C_ILCDOCrecords/ilcdoc-6948-1652866857/

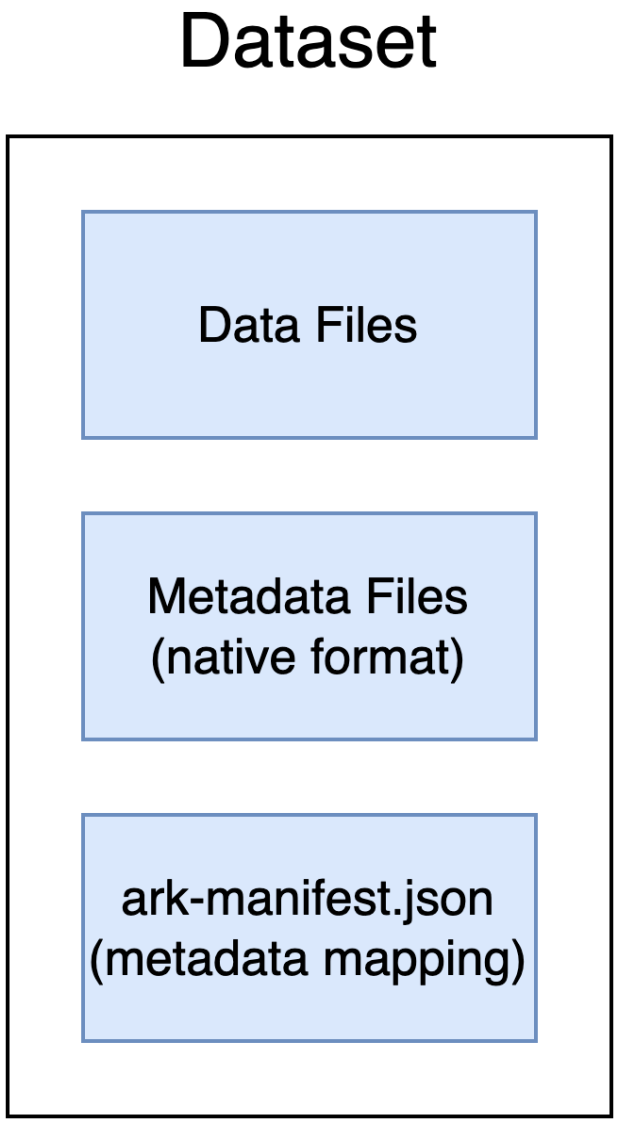
Regards

Arkivum Support

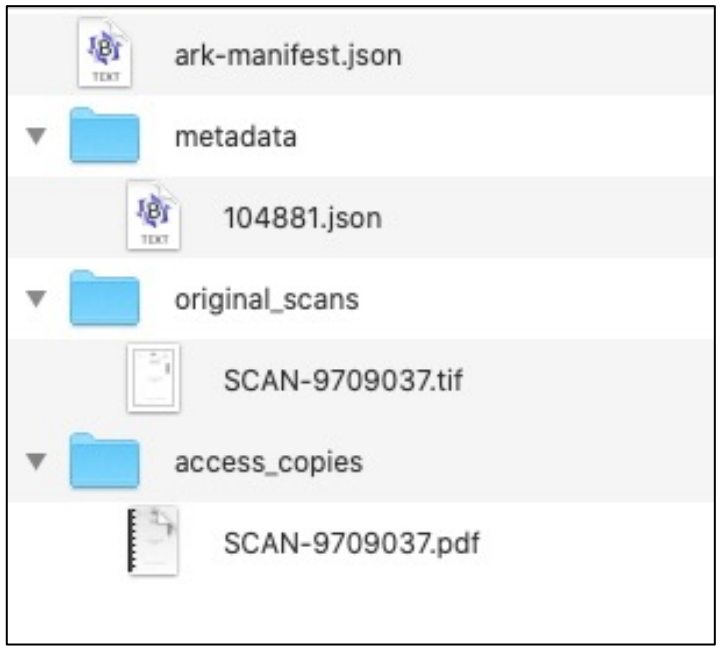
You have been sent this mail because you have an account in Arkivum.com. Please contact us to change your email preferences.

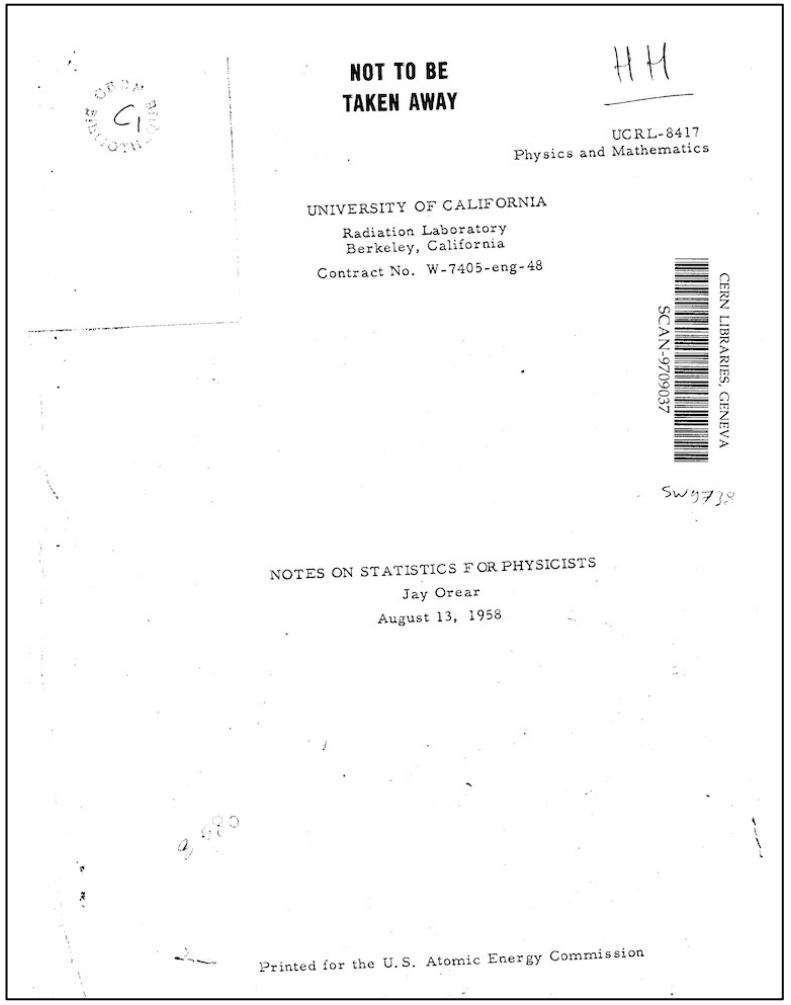
Providing and Updating Metadata

- Metadata can be provided as XML, json or CSV
- Metadata fields are indexed and searchable
- Metadata can be extracted from domain specific metadata files
- Metadata fields can be mapped, e.g. to DublinCore or DataCite
- Rules can be applied to enforce metadata constraints
- Metadata can be provided when data is ingested, or added afterwards
- Metadata can be used to define the structure of datasets
- Metadata can be used to link datasets or files to each other



Metadata Example




NOT TO BE TAKEN AWAY

UCRL-8417
Physics and Mathematics

UNIVERSITY OF CALIFORNIA
Radiation Laboratory
Berkeley, California
Contract No. W-7405-eng-48

CERN LIBRARIES, GENEVA
SCAN-9709037

NOTES ON STATISTICS FOR PHYSICISTS
Jay Orear
August 13, 1958

Printed for the U.S. Atomic Energy Commission

Metadata Example (ark-manifest.json)

Identifier

Type of entity (Collection, Object, File)

Dataset Structure (Parent - Child)

Location of Files (e.g. with ingest)

```
{
    "entities": [
        {
            "id": "104881",
            "entityType": "C",
            "fieldBindings": {
                "metadata/104881.json": [
                    {
                        "key": "dc.title",
                        "valuePathExpression": "$[0]['title']['title']"
                    },
                    {
                        "key": "dc.subject",
                        "valuePathExpression": "$[0]['subject']['term']"
                    }
                ]
            }
        },
        {
            "id": "104881-metadata",
            "entityType": "O",
            "parentIds": [
                "104881"
            ]
        },
        {
            "id": "104881-access-data",
            "entityType": "O",
            "parentIds": [
                "104881"
            ]
        },
        {
            "id": "104881-original-data",
            "entityType": "O",
            "parentIds": [
                "104881"
            ]
        },
        {
            "id": "metadata/104881.json",
            "entityType": "F",
            "uri": "file:///metadata/104881.json",
            "parentIds": [
                "104881-metadata"
            ]
        },
        {
            "id": "access_copies/SCAN-9709037.pdf",
            "entityType": "F",
            "uri": "file:///access_copies/SCAN-9709037.pdf",
            "parentIds": [
                "104881-access-data"
            ],
```

Metadata Example (ark-manifest.json)

File containing native metadata

Mapping to DublinCore

Json Path metadata extraction

Location of native metadata file

```
{
    "entities": [
        {
            "id": "104881",
            "entityType": "C",
            "fieldBindings": {
                "metadata/104881.json": [
                    {
                        "key": "dc.title",
                        "valuePathExpression": "$[0]['title']['title']"
                    },
                    {
                        "key": "dc.subject",
                        "valuePathExpression": "$[0]['subject']['term']"
                    }
                ]
            }
        },
        {
            "id": "104881-metadata",
            "entityType": "O",
            "parentIds": [
                "104881"
            ]
        },
        {
            "id": "104881-access-data",
            "entityType": "O",
            "parentIds": [
                "104881"
            ]
        },
        {
            "id": "104881-original-data",
            "entityType": "O",
            "parentIds": [
                "104881"
            ]
        },
        {
            "id": "metadata/104881.json",
            "entityType": "F",
            "uri": "file:///metadata/104881.json",
            "parentIds": [
                "104881-metadata"
            ]
        },
        {
            "id": "access_copies/SCAN-9709037.pdf",
            "entityType": "F",
            "uri": "file:///access_copies/SCAN-9709037.pdf",
            "parentIds": [
                "104881-access-data"
            ],
```

Metadata Example (ark-manifest.json)

Where file should be in dataset hierarchy

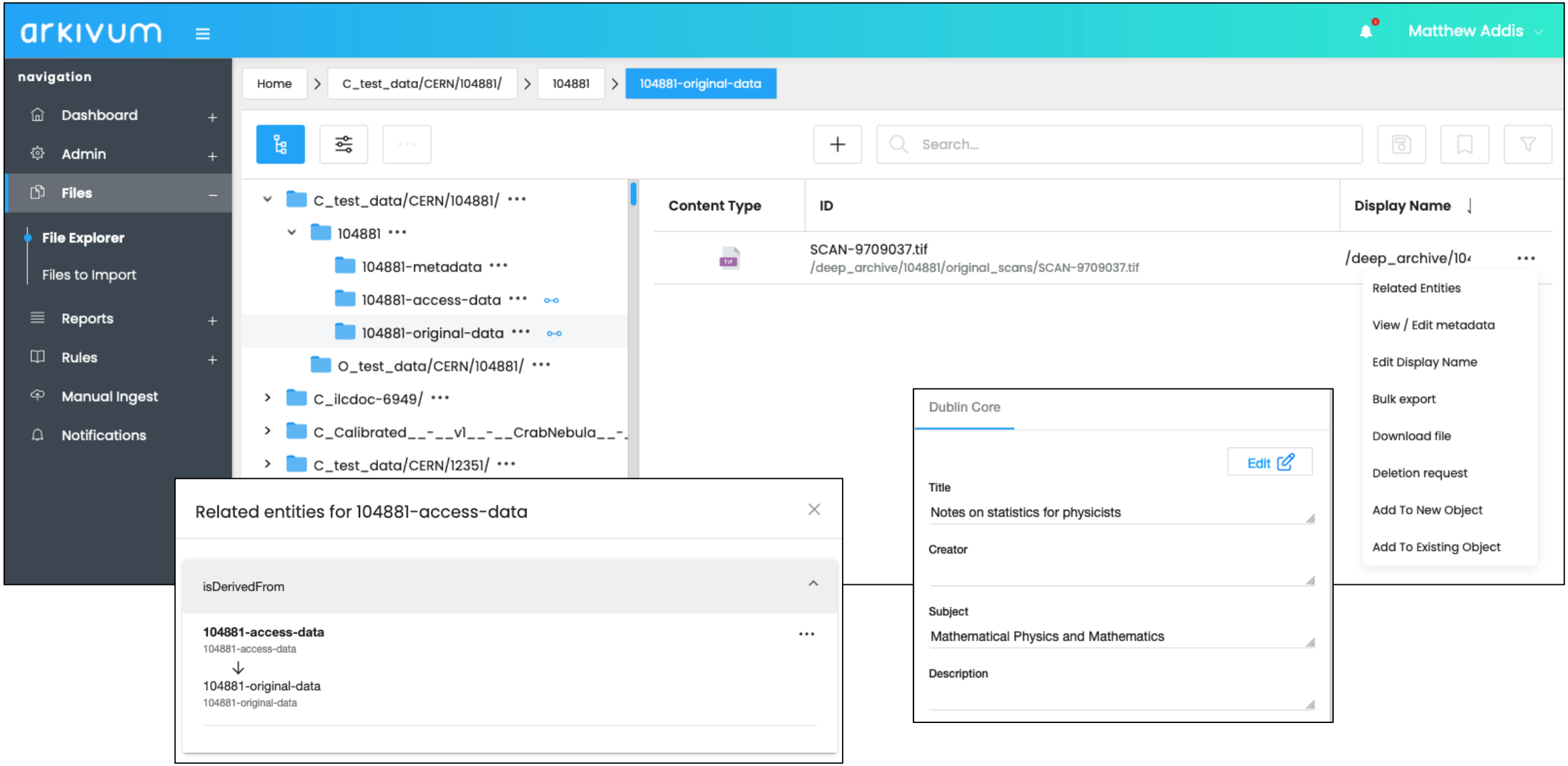
Relationship to other entities

Type of relationship (DataCite relations)

Where to store the file

```
{
    "id": "access_copies/SCAN-9709037.pdf",
    "entityType": "F",
    "uri": "file:///access_copies/SCAN-9709037.pdf",
    "parentIds": [
        "104881-access-data"
    ],
    "fieldBindings": {
        "metadata/104881.json": [
            {
                "key": "dc.format",
                "valuePathExpression": "$[0]['files'][?(@['full_name']=='SCAN-9709037.pdf')]['eformat']"
            },
            {
                "key": "dc.title",
                "valuePathExpression": "$[0]['files'][?(@['full_name']=='SCAN-9709037.pdf')]['full_name']"
            },
            {
                "key": "dc.type",
                "valuePathExpression": "$[0]['files'][?(@['full_name']=='SCAN-9709037.pdf')]['type']"
            },
            {
                "key": "dc.source",
                "valuePathExpression": "$[0]['files'][?(@['full_name']=='SCAN-9709037.pdf')]['url']"
            }
        ]
    },
    "relation": {
        "id": "original_scans/SCAN-9709037.tif",
        "kind": "isDerivedFrom"
    }
},
{
    "id": "original_scans/SCAN-9709037.tif",
    "entityType": "F",
    "uri": "file:///original_scans/SCAN-9709037.tif",
    "parentIds": [
        "104881-original-data"
    ],
    "fieldBindings": {
        "metadata/104881.json": [
            {
                "key": "dc.format",
                "valuePathExpression": "$[0]['files'][?(@['full_name']=='SCAN-9709037.tif')]['eformat']"
            },
            {
                "key": "dc.title",
                "valuePathExpression": "$[0]['files'][?(@['full_name']=='SCAN-9709037.tif')]['full_name']"
            },
            {
                "key": "dc.type",
                "valuePathExpression": "$[0]['files'][?(@['full_name']=='SCAN-9709037.tif')]['type']"
            },
            {
                "key": "dc.source",
                "valuePathExpression": "$[0]['files'][?(@['full_name']=='SCAN-9709037.tif')]['url']"
            }
        ]
    },
    "datapool": "Deep Archive"
}
```

End Result

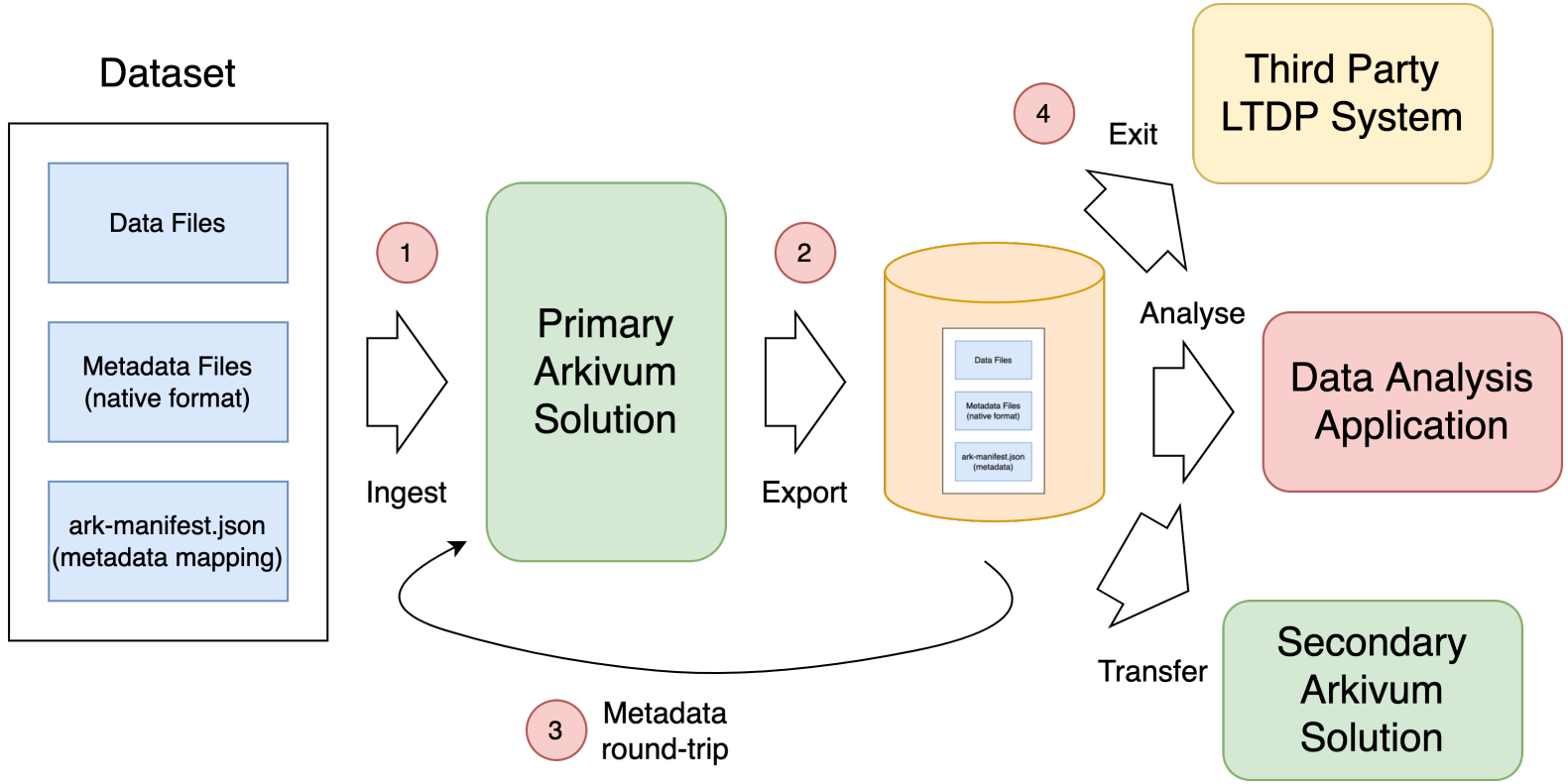

arkivum
Matthew Addis
navigation
Dashboard
Admin
Files
File Explorer
Files to Import
Reports
Rules
Manual Ingest
Notifications
Home
C_test_data/CERN/104881/
104881
104881-original-data
Search...
C_test_data/CERN/104881/
104881
104881-metadata
104881-access-data
104881-original-data
O_test_data/CERN/104881/
C_ilcdoc-6949/
C_Calibrated__-__v1__-__CrabNebula__-_
C_test_data/CERN/12351/
Content Type
ID
Display Name
SCAN-9709037.tif
/deep_archive/104881/original_scans/SCAN-9709037.tif
/deep_archive/10
Related Entities
View / Edit metadata
Edit Display Name
Bulk export
Download file
Deletion request
Add To New Object
Add To Existing Object
Related entities for 104881-access-data
isDerivedFrom
104881-access-data
104881-access-data
104881-original-data
104881-original-data
Dublin Core
Edit
Title
Notes on statistics for physicists
Creator
Subject
Mathematical Physics and Mathematics
Description

Metadata Updates

- Add metadata to existing datasets using a metadata file, e.g. ark-manifest.json
- Export metadata file, update file, and re-ingest updated metadata file
- GET/POST/PUT metadata using REST API
- Edit metadata in the UI
- Extract metadata fields from native metadata files already in the archive
- Trigger external scripts/applications that extract/generate metadata using webhooks, Cloud Events, Arkivum REST API

Round-trip, dataset migration, exit strategy

- Export datasets into a bucket
- Bagit for dataset integrity checks
- Original data files, metadata files
- Normalised preservation versions of files (if created)
- ark-manifest.json description

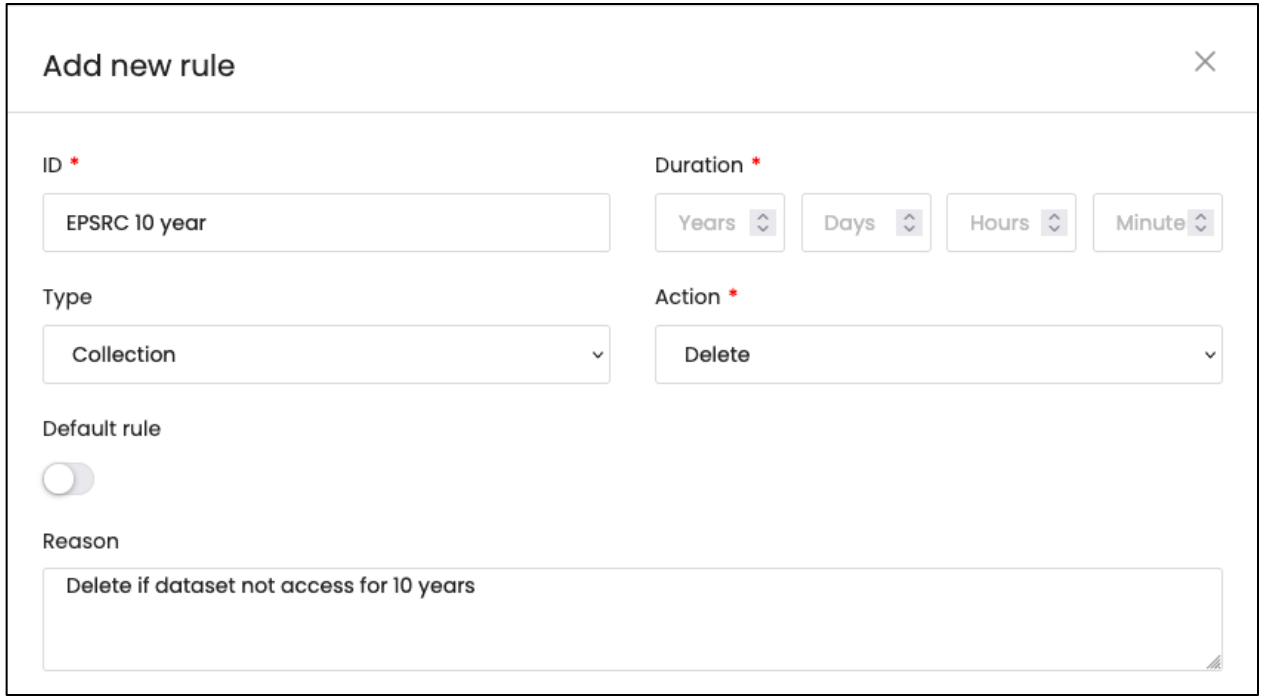


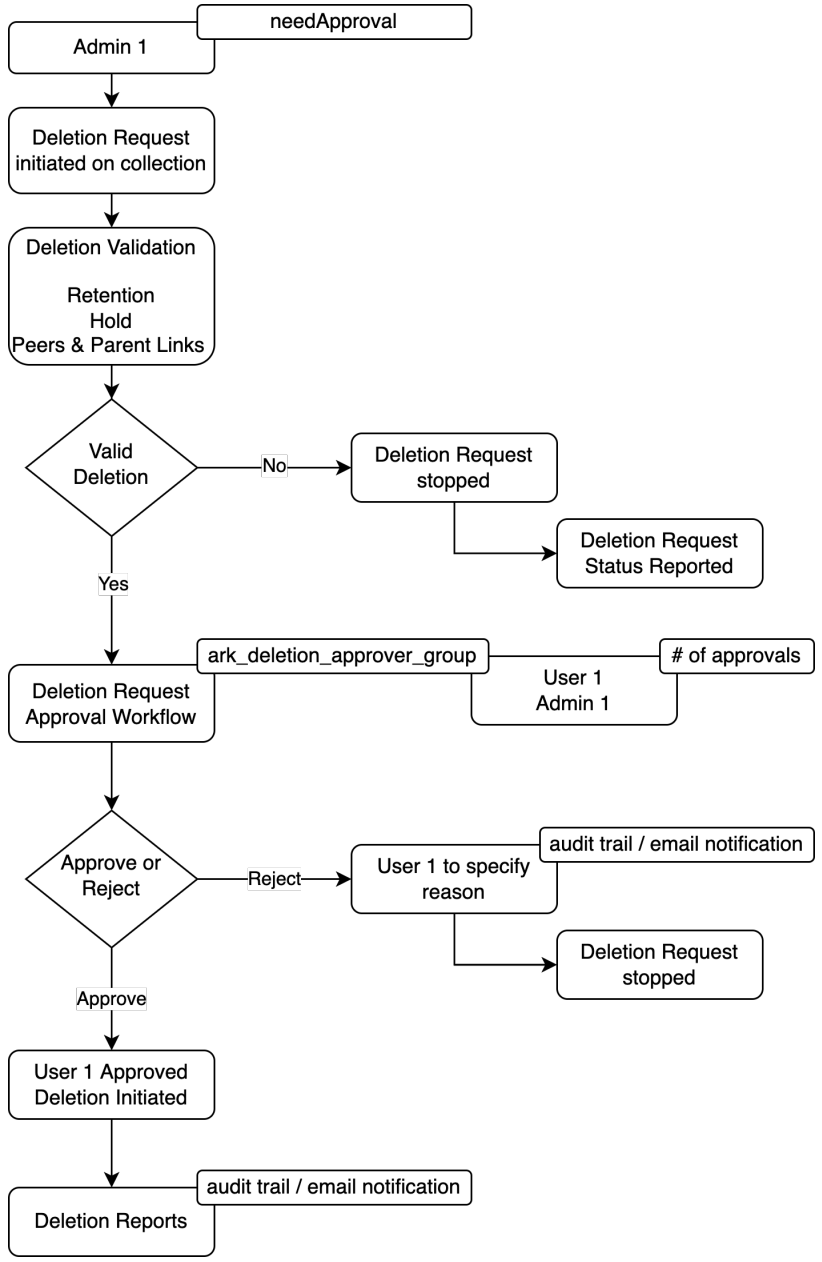
Round trip dataset

https://vimeo.com/723624529/0c4b8419e1

Retention Schedules and Deletion

- Dataset deletion follows a request/approve workflow
- Deletion needs one approvals
- Retention schedules can create deletion events

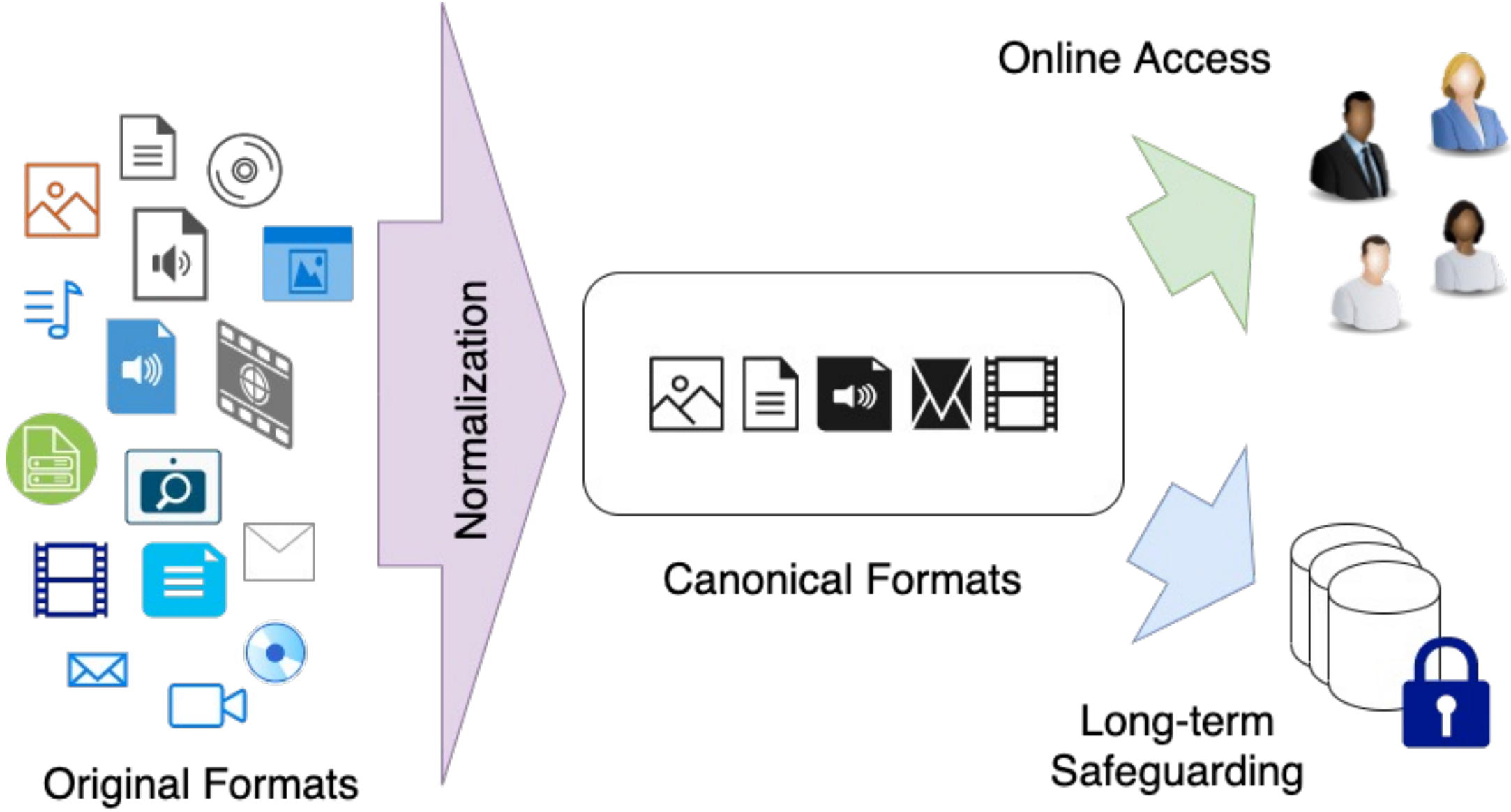




Solution: Part 3

Agenda

- File Format Normalisation
- Long-term storage: replication, tiers, fixity checks
- Access to archived data using xrootd
- Integration with InvenioRDM
- Co-locating applications with the Arkivum solution for pre and post processing of archived data
- Summary


Online Access
Normalization
Canonical Formats
Original Formats
Long-term
Safeguarding

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

Preferred Formats

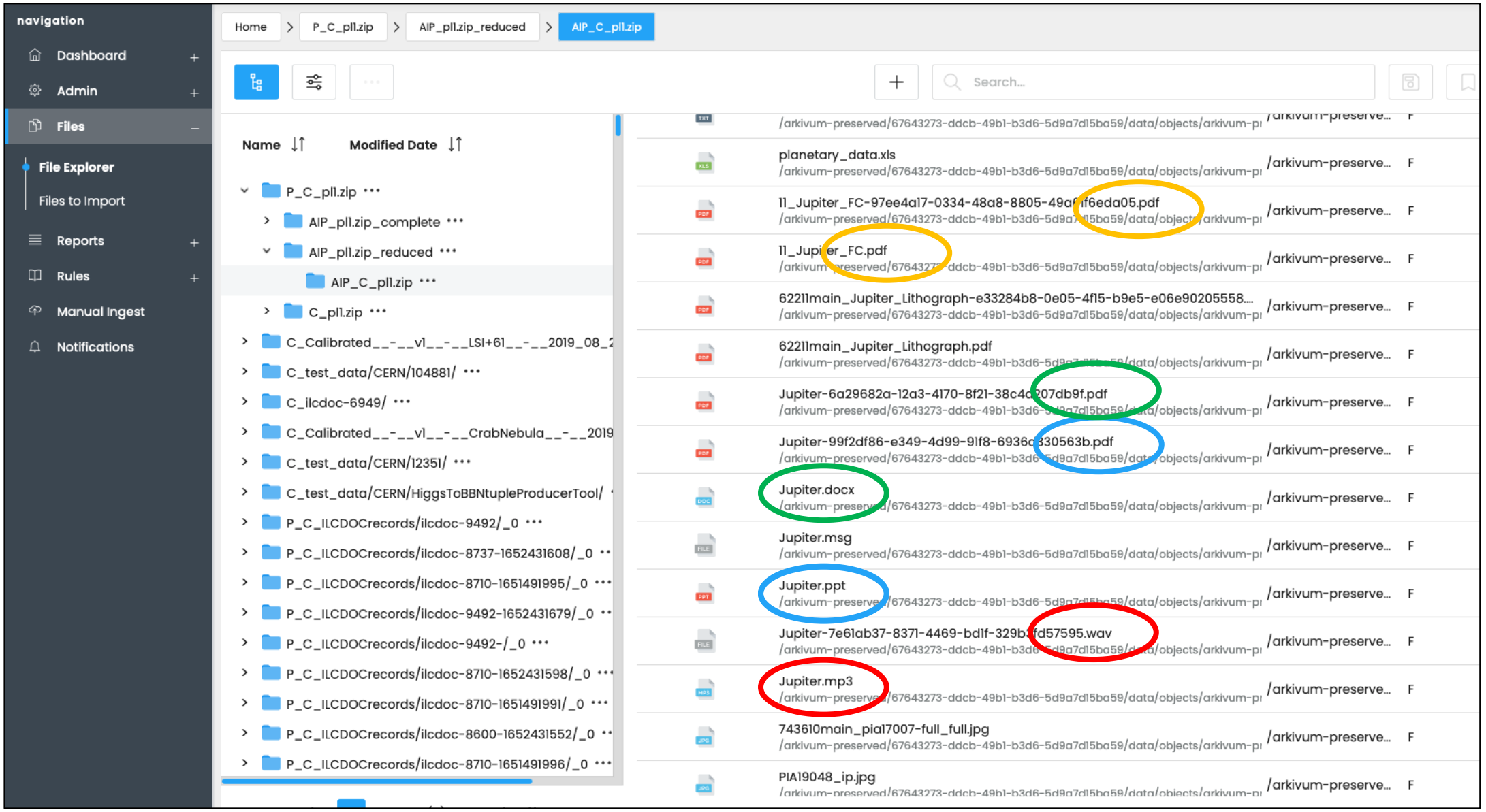
Preferred Formats	Format Versions	Format Specifications
ASCII Text	7 bit	ISO/IEC 646:1991 Information technology -- ISO 7-bit coded character set for information interchange: (http://www.iso.org/iso/catalogue_detail.htm?csnumber=4777)
Unicode Text	UTF-8	RTF 3629: UTF-8, A Transformation Format of ISO 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)

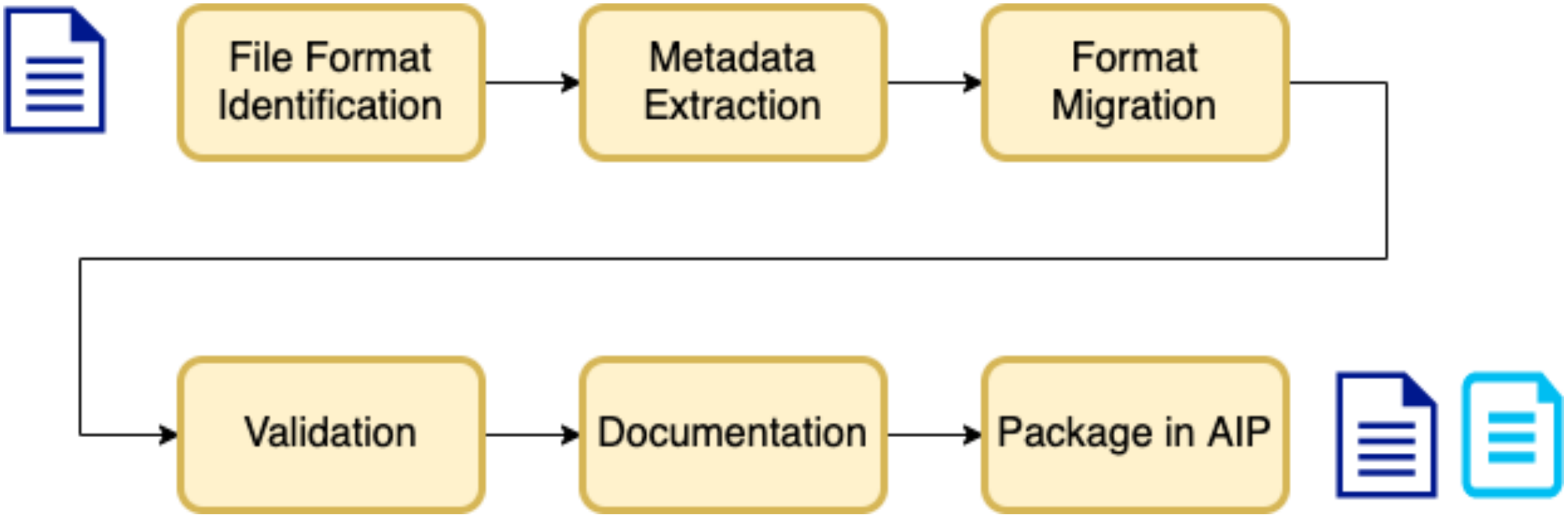
Acceptable Formats

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)

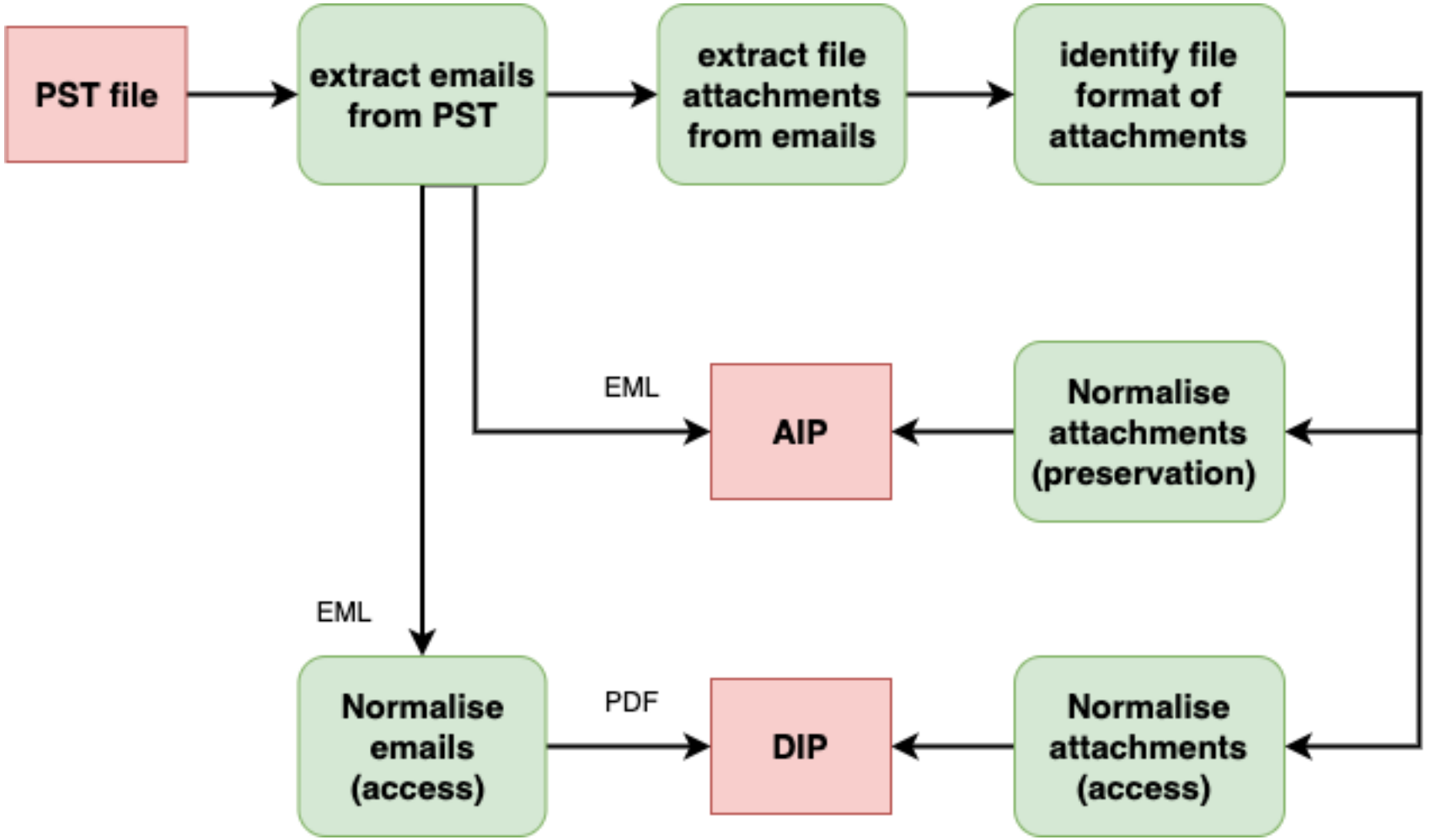
https://www.archives.gov/records-mgmt/policy/transfer-guidance-tables.html

File Format Normalisation

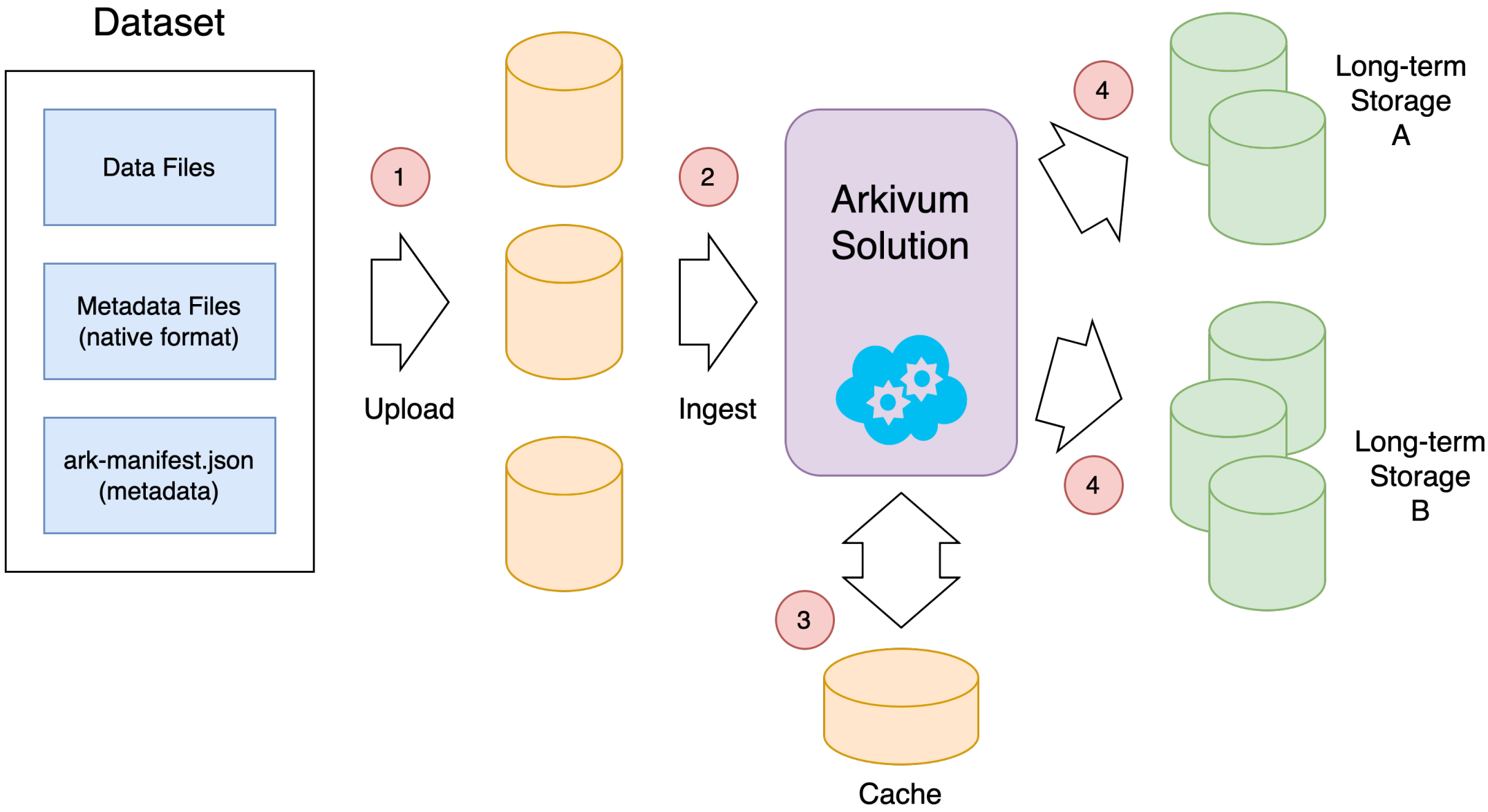



File Format Identification
Metadata Extraction
Format Migration
Validation
Documentation
Package in AIP

Email Normalization Workflow

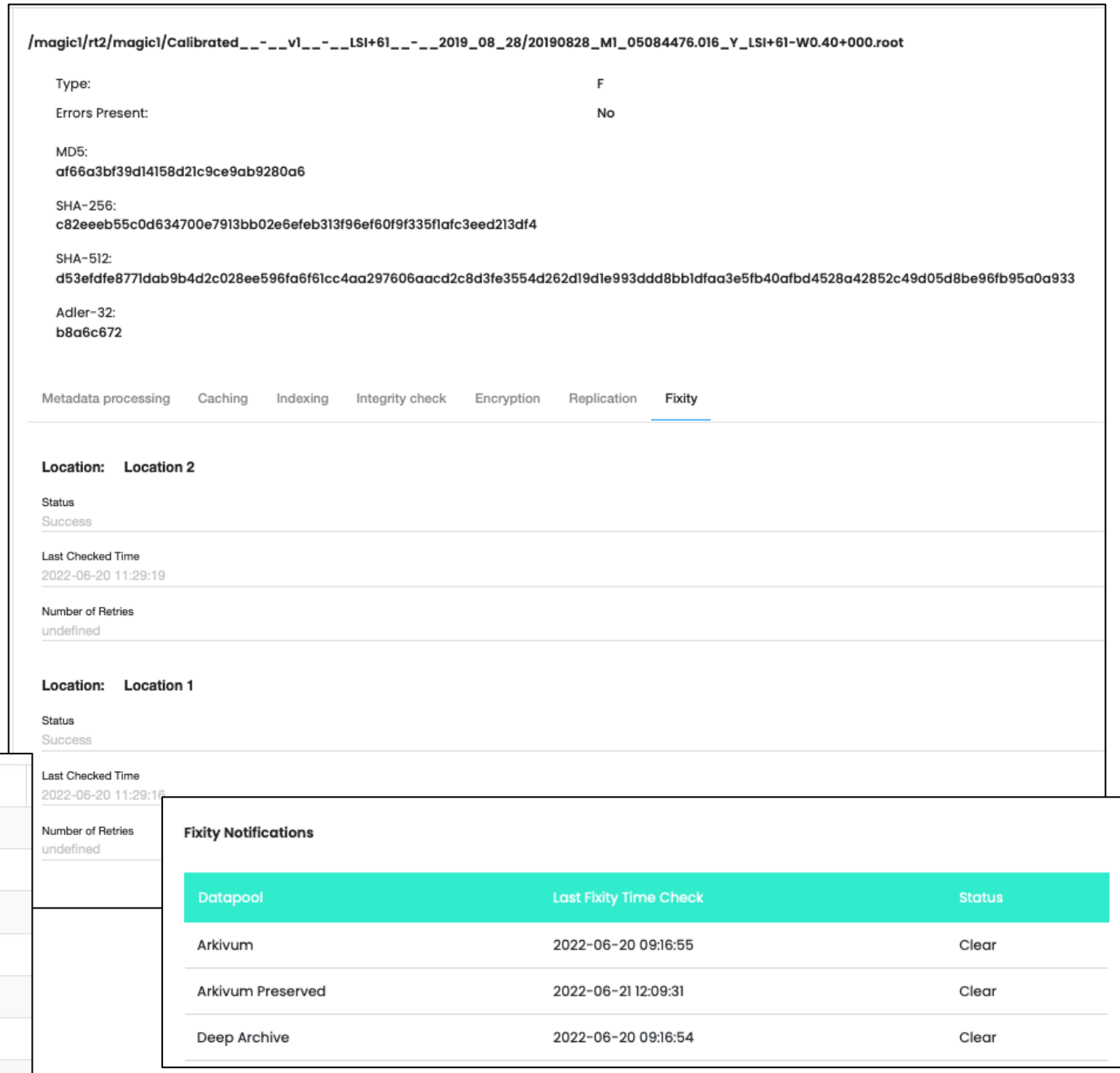


Long-term Archival Storage in Buckets



Storage Options and Fixity Checks

- Multiple copies of each file
- Multiple cloud providers
- Combine different tiers of storage
- Multiple checksums for each file
- Data integrity checks
  - When data first stored
  - Periodic checks

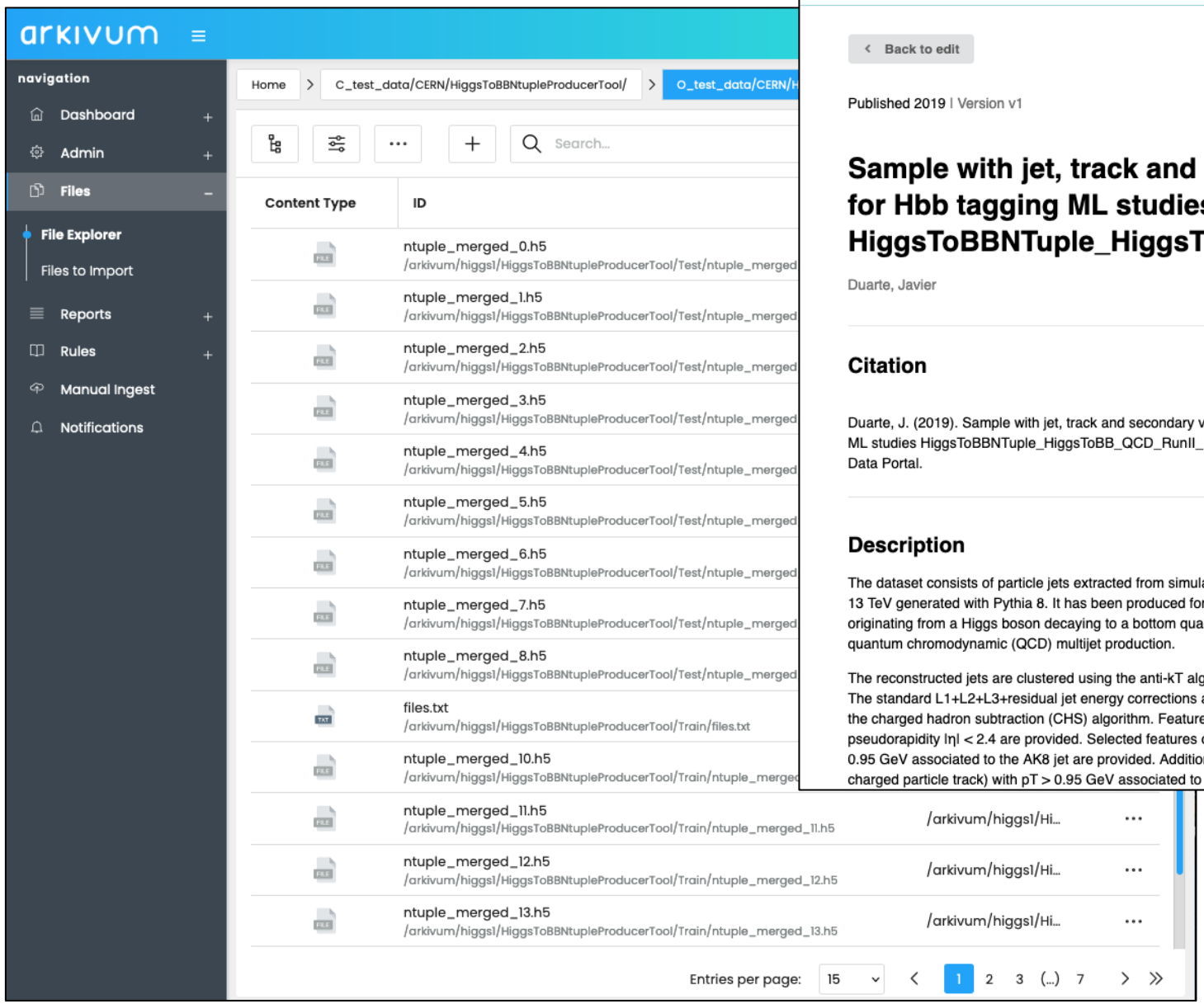


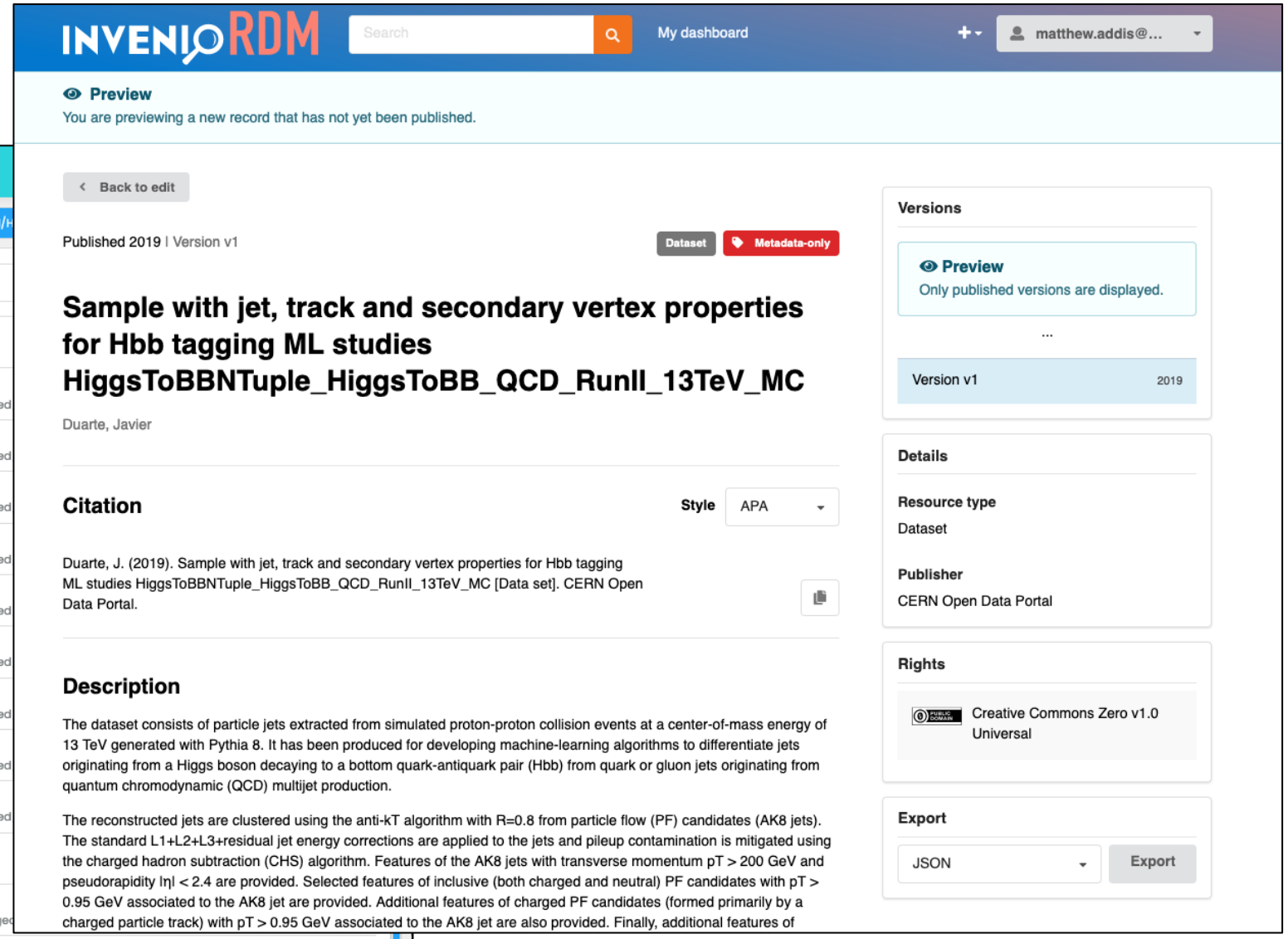
Event Type Code	Event	ID	Action	Time Stamp
OBJECT_FIXITY_CHECK_SUCCESS	Fixity	62b1b47106d6c474d83b968d	Fixity performed	2022-06-21 12:09:31
OBJECT_FIXITY_CHECK_SUCCESS	Fixity	62b1b47106d6c474d83b9689	Fixity performed	2022-06-21 12:09:31
OBJECT_FIXITY_CHECK_SUCCESS	Fixity	62b1b47106d6c474d83b969b	Fixity performed	2022-06-21 12:09:31
OBJECT_FIXITY_CHECK_SUCCESS	Fixity	62b1b47106d6c474d83b9689	Fixity performed	2022-06-21 12:09:31
OBJECT_FIXITY_CHECK_SUCCESS	Fixity	62b1b47106d6c474d83b969b	Fixity performed	2022-06-21 12:09:31
OBJECT_FIXITY_CHECK_SUCCESS	Fixity	62b1b47306d6c474d83b96ae	Fixity performed	2022-06-21 12:09:31
OBJECT_FIXITY_CHECK_SUCCESS	Fixity	62b1b47106d6c474d83b9695	Fixity performed	2022-06-21 12:09:31

Datapools and Storage

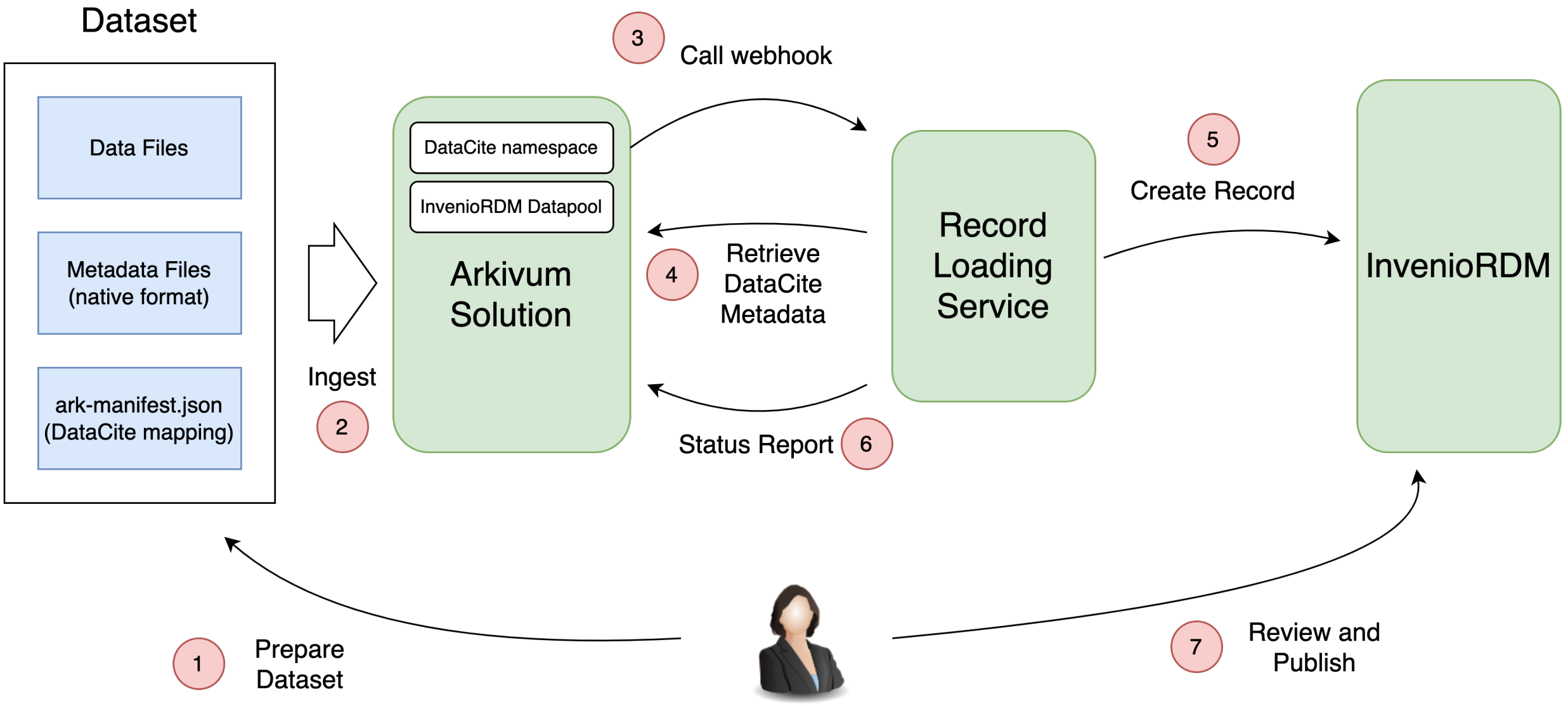
https://vimeo.com/722930836/9659bb1784

Publication, Discovery, Analysis and Reuse of Research Data


arkivum
navigation
Dashboard
Admin
Files
File Explorer
Files to Import
Reports
Rules
Manual Ingest
Notifications
Home
C_test_data/CERN/HiggsToBBNtupleProducerTool/
O_test_data/CERN/H
Search...
Content Type
ID
ntuple_merged_0.h5
/arkivum/higgs1/HiggsToBBNtupleProducerTool/Test/ntuple_merged
ntuple_merged_1.h5
/arkivum/higgs1/HiggsToBBNtupleProducerTool/Test/ntuple_merged
ntuple_merged_2.h5
/arkivum/higgs1/HiggsToBBNtupleProducerTool/Test/ntuple_merged
ntuple_merged_3.h5
/arkivum/higgs1/HiggsToBBNtupleProducerTool/Test/ntuple_merged
ntuple_merged_4.h5
/arkivum/higgs1/HiggsToBBNtupleProducerTool/Test/ntuple_merged
ntuple_merged_5.h5
/arkivum/higgs1/HiggsToBBNtupleProducerTool/Test/ntuple_merged
ntuple_merged_6.h5
/arkivum/higgs1/HiggsToBBNtupleProducerTool/Test/ntuple_merged
ntuple_merged_7.h5
/arkivum/higgs1/HiggsToBBNtupleProducerTool/Test/ntuple_merged
ntuple_merged_8.h5
/arkivum/higgs1/HiggsToBBNtupleProducerTool/Test/ntuple_merged
files.txt
/arkivum/higgs1/HiggsToBBNtupleProducerTool/Train/files.txt
ntuple_merged_10.h5
/arkivum/higgs1/HiggsToBBNtupleProducerTool/Train/ntuple_merged
ntuple_merged_11.h5
/arkivum/higgs1/HiggsToBBNtupleProducerTool/Train/ntuple_merged_11.h5
/arkivum/higgs1/Hi...
ntuple_merged_12.h5
/arkivum/higgs1/HiggsToBBNtupleProducerTool/Train/ntuple_merged_12.h5
/arkivum/higgs1/Hi...
ntuple_merged_13.h5
/arkivum/higgs1/HiggsToBBNtupleProducerTool/Train/ntuple_merged_13.h5
/arkivum/higgs1/Hi...
Entries per page:
15
1
2
3
(...)
7


INVENIO RDM
Search
My dashboard
matthew.addis@...
Preview
You are previewing a new record that has not yet been published.
Back to edit
Published 2019 | Version v1
Dataset
Metadata-only
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
JSON
Export

InvenioRDM Workflow

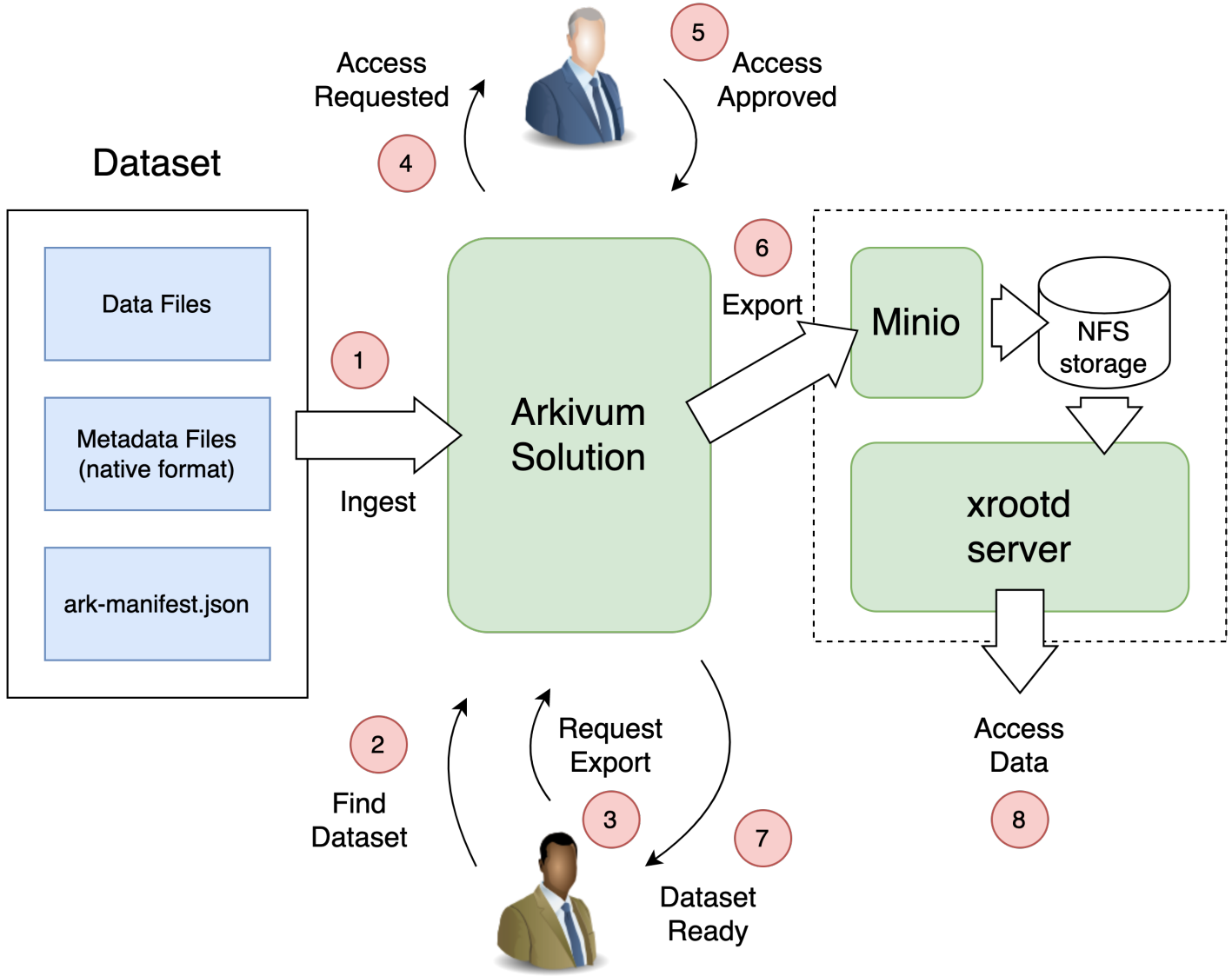


InvenioRDM Workflow

https://vimeo.com/723439662/260a70b326

Xrootd Workflow

- Xrootd server hosted in GCP
- Transfer of datasets on-demand

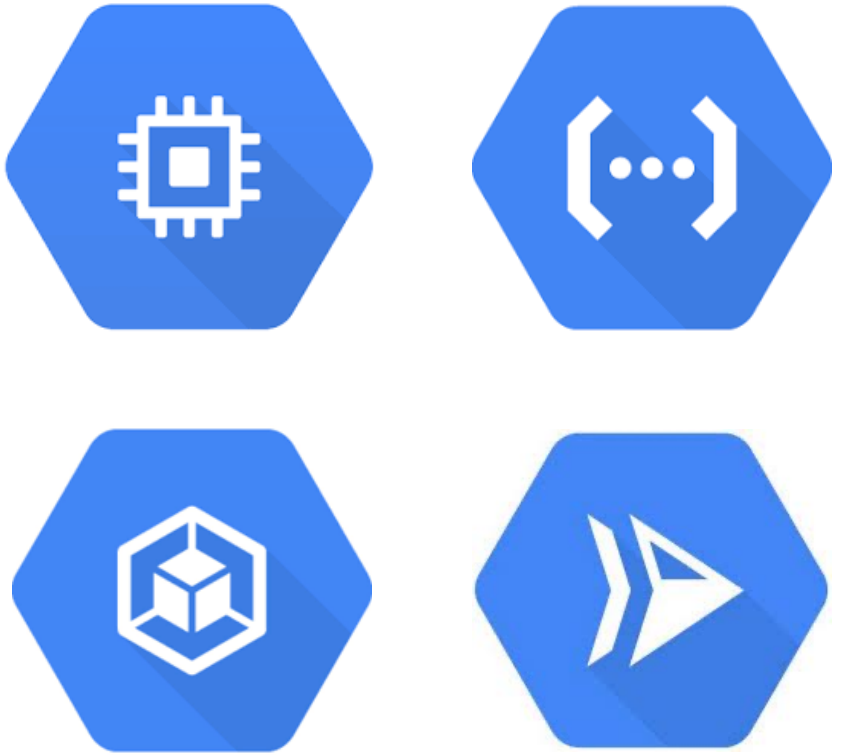


Xrootd Workflow

https://vimeo.com/723493247/212b1f4de5

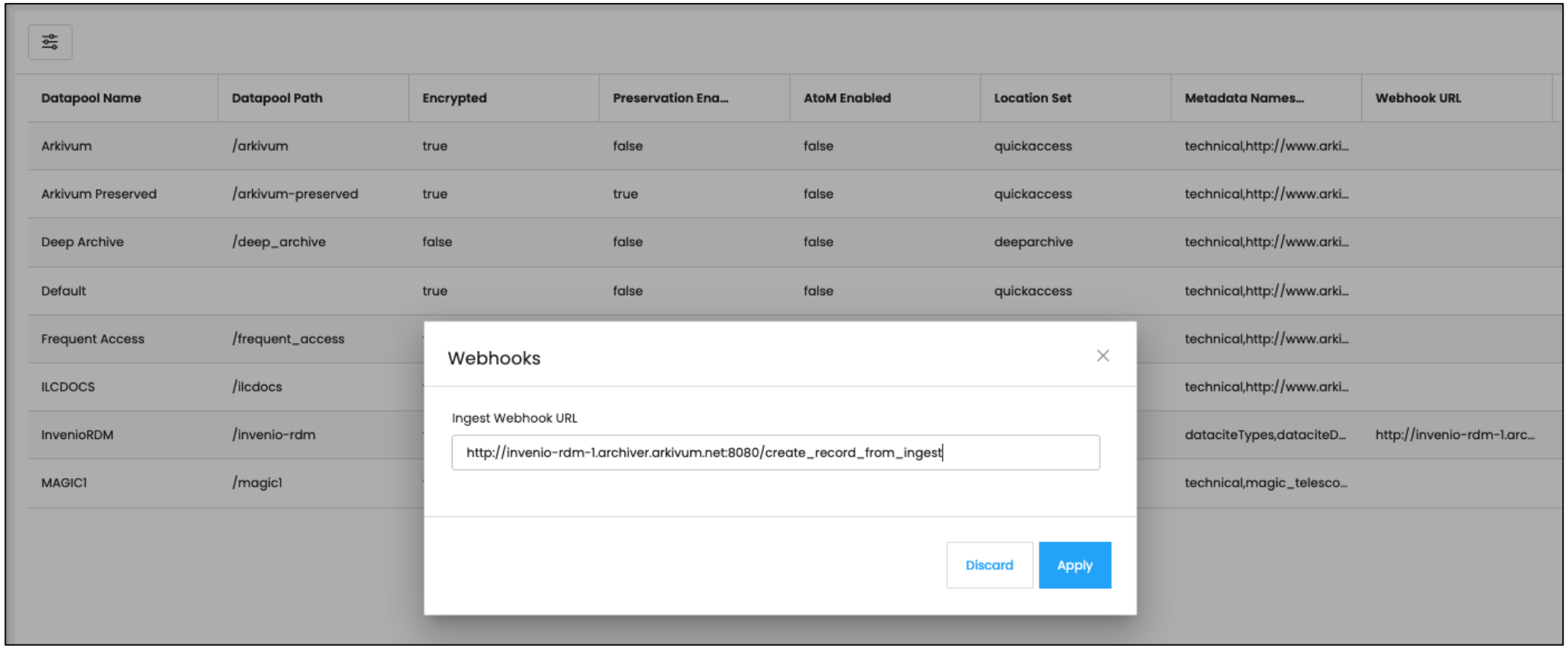
Cloud-hosted Processing of Archived Research Datasets

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



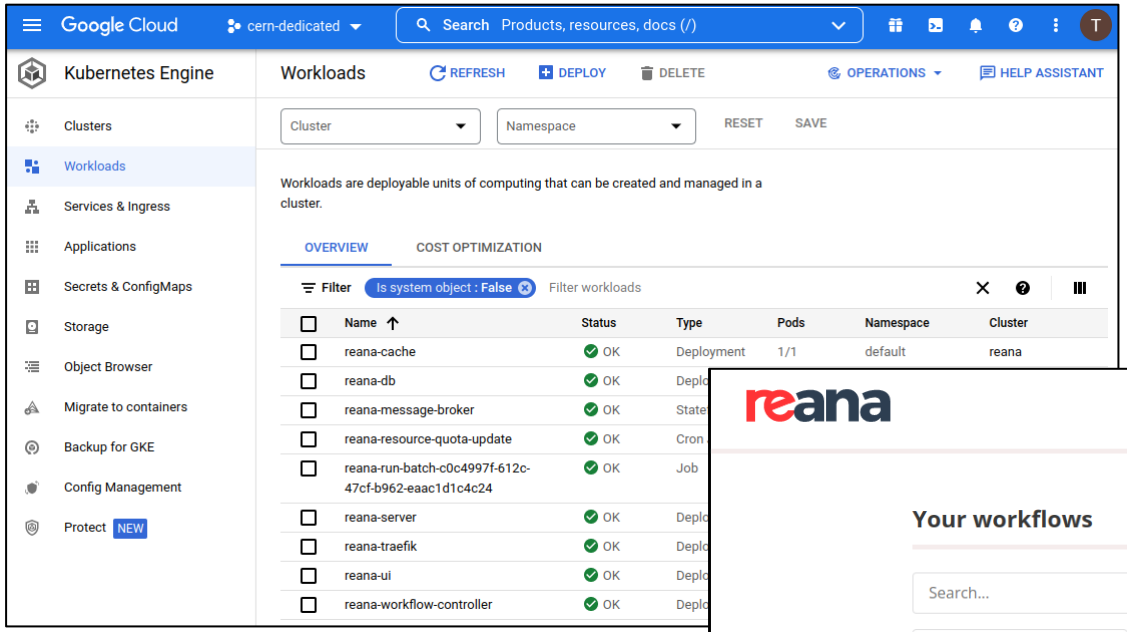
Webhooks

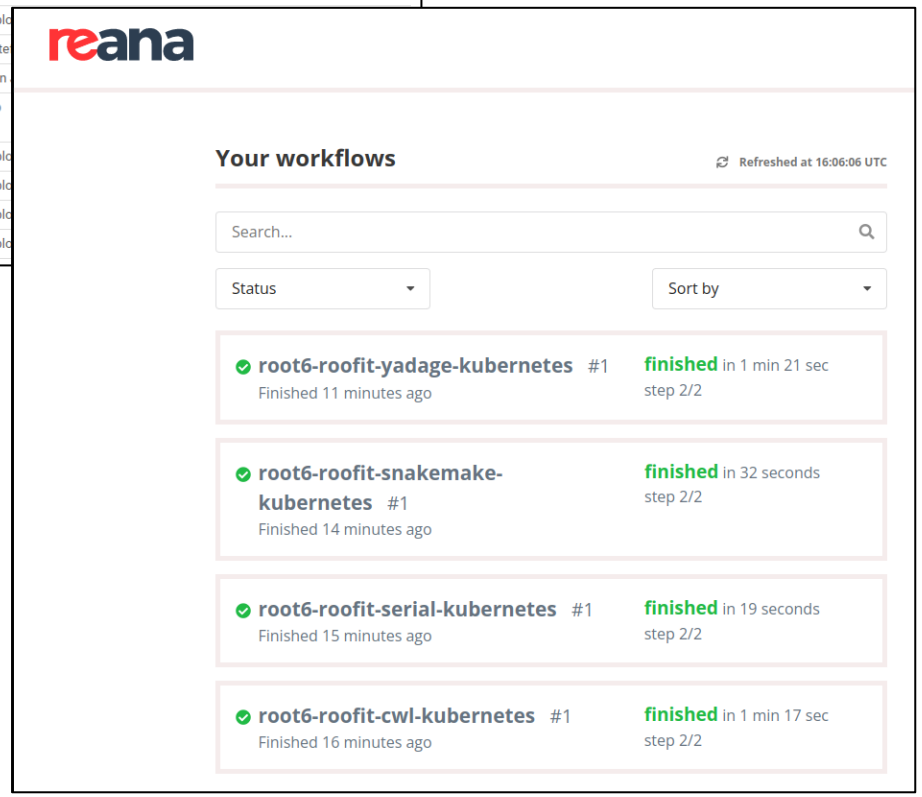
- Can be set per job (ingest, preservation, export)
- Default webhooks for a Datapool (ingest, preservation)
- Includes API tokens to allow calls to secured endpoints

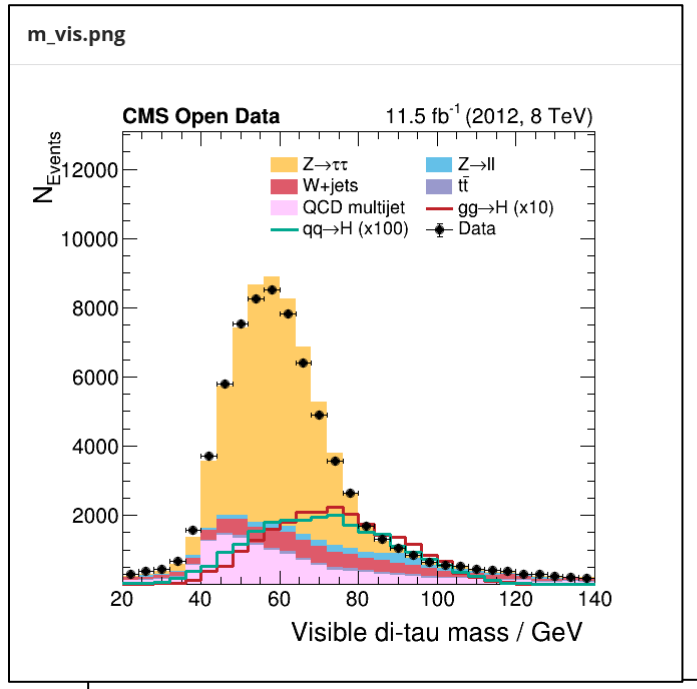


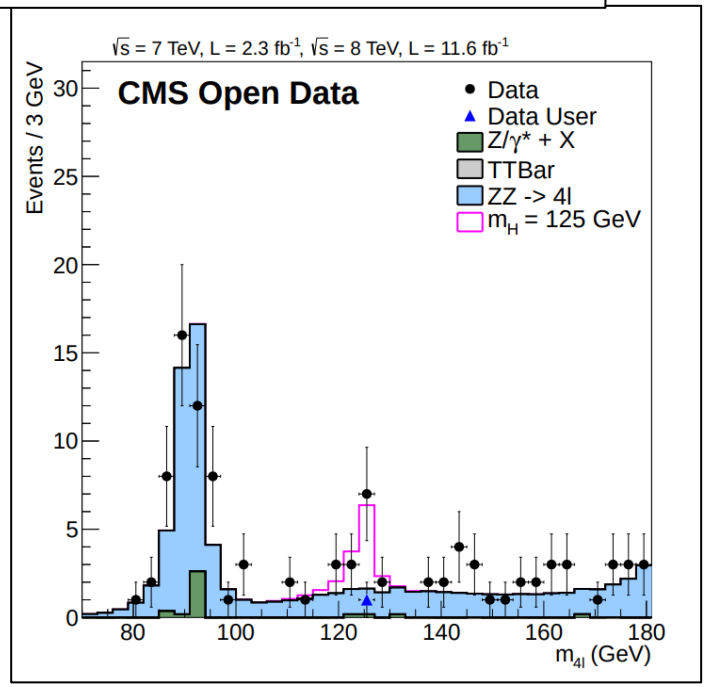
reana

snakemake









Long-Term Digital Preservation of Research Data

TRUST

RDA
RESEARCH DATA ALLIANCE

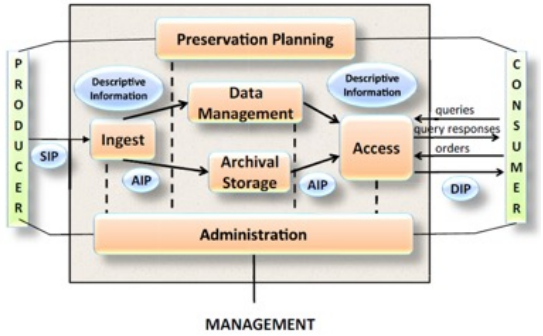
CORE TRUST SEAL

DigitalPreservationCoalition

FAIRSFAIR
Fostering Fair Data Practices in Europe

NDSA

Levels of Digital Preservation



F	R13	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**
A	R10 R15 R16	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**
I	R14	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**
R	R2 R7 R14	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**

Self-Assessment using the CTS+FAIR Capability Maturity Model

Area	Short Requirement Name	FAIR Principle (Nature)	RDA Indicator	FAIRsFAIR Metric	Evidence Links
Digital Object Management	**13. Data discovery and identification**	**Discovery** F2. data are described with rich metadata (defined by R1 below)	RDA-F2-01M Rich metadata is provided to allow discovery (Essential)	FsF-F2-01M Metadata includes descriptive core elements (creator, title, data identifier, publisher, publication date, summary and keywords) to support data findability.	Metadata support in the Arkivum solution includes DublinCore, DataCite and custom fields. Metadata can be ingested and stored in its native format (json and XML), e.g. domain specific schemas. Metadata fields in domain specific metadata can be mapped to DC or DataCite and then made searchable. DataCite metadata can be exported to InvenioRDM.
		Discovery F4. (meta)data are registered or indexed in a searchable resource	RDA-F4-01M Metadata is offered in such a way that it can be harvested and indexed (Essential)	FsF-F4-01M Metadata is offered in such a way that it can be retrieved by machines.	InvenioRDM supports publication of metadata landing pages. InvenioRDM supports metadata harvesting using OAI-PMH. https://inveniordm.docs.cern.ch/reference/oai_pmh/ InvenioRDM supports PID services, e.g. DOI registration using DataCite. https://inveniordm.docs.cern.ch/customize/dois/ Metadata in the Arkivum solution can be searched and retrieved though an API (authenticated users only)
		Identification F1. (meta)data are assigned a globally unique and persistent identifier	RDA-F1-01M Metadata is identified by a persistent identifier (Essential) RDA-F1-01D Data is identified by a persistent identifier (Essential) RDA-F1-02M Metadata is identified by a globally unique identifier (Essential) RDA-F1-02D Data is identified by a globally unique identifier (Essential)	FsF-F1-02D Data is assigned a persistent identifier. FsF-F1-01D Data is assigned a globally unique identifier.	All entities in the Arkivum solution (PCDM Collections, objects, files) are assigned globally unique UUIDs. Users can provide one or more of their own identifiers as metadata, e.g. PIDs such as DOIs or handles. The Arkivum metadata schema supports identifier type and indentifier value. URLs are possible to entities in the Arkivum solution (authentication is required for access). Records in InvenioRDM can have one or more PIDs, including those registered with external PID systems or as Alternative Identifiers. Identifier schemes include: ARK, arXiv, Bibcode, DOI, EAN13, EISSN, Handle, IGSN, ISBN, ISSN, ISTC, LISSN, LSID, PubMed ID, PURL, UPC, URL, URN, W3ID. https://inveniordm.docs.cern.ch/reference/metadata/#external-pids https://inveniordm.docs.cern.ch/reference/metadata/#alternate-identifiers-0-n
		Identification F3. metadata clearly and explicitly include the identifier of the data it describes	RDA-F3-01M Metadata includes the identifier for the data (Essential)	FsF-F3-01M Metadata includes the identifier of the data it describes.	Metadata in InvenioRDM includes identifiers (see above). Likewise, all records in the Arkivum solution have identfiers. Records in InvenioRDM can include files. The InvenioRDM roadmap includes support for referencing data in external systems. https://inveniosoftware.org/products/rdm/roadmap/

Further Materials and Summary

Further Materials

- User Guide
- API Guide
- Cookbooks and examples for specific activities
- Demos and pilot projects
- Onboarding process, including detailed training
- Zendesk support portal
- Mappings to NDSA preservation levels, DPC RAM, CTS, FAIR metrics
- iPRES 2021 paper and iPRES 2022 panel
- ARCHIVER deliverables

Summary

- The Arkivum solution enables LTDP of research data:
  - Ingest, Preserve, Safeguard, Search, Navigate, Download, Export, Publish
- Support for reuse of scientific datasets:
  - xrootd, InvenioRDM, webhooks, reana, snakemake, scripts, k8s, VMs
- Flexible deployment
  - GCP, AWS, Open Stack
- Authentication and Authorisation
  - eduGAIN, request/approve workflows, control over actions and data
- Support for good practice, assessment and certification
  - NDSA preservation levels, DPC RAM, CoreTrustSeal, FAIR principles and metrics
- Available for use today
  - Listed on EOSC Marketplace, contact Arkivum or Google

NDSA

Levels of Digital Preservation

CORE TRUST SEAL

FAIRSFAIR

arkivum
Bringing archived data to life

QUESTIONS?

matthew.addis@arkivum.com

orcid.org/0000-0002-3837-2526

www.arkivum.com