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

Data Sources

Data supported includes:
- Arkivum Fully Managed Archiving and Preservation Solution
- Arkivum Archiving & Preservation Solution
- Archivists, Researchers, Compliance etc.

Safeguarding
Preservation
Discovery & Access

Data Uploaded to Arkivum

Managed Ingest Process

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
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
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

Reperform the Higgs discovery analysis on 70 TB of CMS open data in a live demo

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

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

Layer 2
Preservation
- OAIS conformant services: data readability formats, normalization, obsolescence monitoring, files fixity, authenticity checks, etc.
  - ISO 14721/16363, 26324 and related standards

Layer 3
Baseline user services
- User services: search, discover, share, indexing, data removal, etc.
  - Access under Federated IAM

Layer 4
Advanced services
- High level services: visual representation of data (domain specific), reproducibility of scientific analyses, etc.

Scientific use cases deployments: [https://www.archiver-project.eu/deployment-scenarios](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](https://doi.org/10.5281/zenodo.3618215)
Long-Term Digital Preservation of Research Data: Good Practice

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
R1. (meta)data are retrievable by their identifier using a standardized communications protocol.
A1.1 the protocol is open, free, and universally implementable (vs context)
A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
A2. Security
R10. Preservation plan
R14. ReUse
R1. (meta)data are richly described with a plurality of accurate and relevant attributes.

R7. Data integrity and authenticity
R1.2. (meta)data are associated with detailed provenance.
R13. (meta)data meet domain-relevant community standards.
Standards, External Assessment and Certification

- Self Assessment
- Peer Review
- Formal Certification
Long-Term Digital Preservation (LTDP) Factories

Content types and sources

Automated Workflows

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

FAIR data for Researchers

API

images: Flaticon.com
Arkivum Microservices and Workflows

Levels of Digital Preservation

[Diagram showing microservices and workflows]

- Dataset checks
- Copy to cache
- Checksum validation
- File format identification
- Virus scanning
- Bagit validation
- Package extraction (zip, tar, 7z)
- Metadata extraction
- Metadata import
- Normalisation (Archivematica)
- Content organisation (PCDM)
- Encryption
- Packaging (Archivematica)
- Metadata indexing
- Full text indexing
- Replication
- Data checks
Levels of Digital Preservation Assessment Tool

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

### Functional Area

<table>
<thead>
<tr>
<th>Functional Area</th>
<th>Level 1 (Know your content)</th>
<th>Level 2 (Protect your content)</th>
<th>Level 3 (Monitor your content)</th>
<th>Level 4 (Sustain your content)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Storage</td>
<td>2 (Have two complete copies in separate locations)</td>
<td>2 (Have three complete copies with at least one copy in a separate geographic location)</td>
<td>2 (Have at least one copy in a geographic location with a different disaster threat than the other copies)</td>
<td>2 (Have at least three copies in geographic locations, each with a different disaster threat)</td>
</tr>
<tr>
<td>Integrity</td>
<td>2 (Verify integrity information if it has been provided with the content)</td>
<td>2 (Verify integrity information when moving or copying content)</td>
<td>2 (Verify integrity information of content at fixed intervals)</td>
<td>2 (Verify integrity information in response to specific events or activities)</td>
</tr>
<tr>
<td>Control</td>
<td>2 (Determine the human and software agents that should be authorized to read, write, move, and delete content)</td>
<td>2 (Document the human and software agents authorized to read, write, and delete content and apply it)</td>
<td>2 (Perform periodic review of actions/access logs)</td>
<td>2 (Perform audit of integrity information on demand)</td>
</tr>
<tr>
<td>Metadata</td>
<td>2 (Create inventory of content, also documenting current storage locations)</td>
<td>2 (Verify file formats and other essential content characteristics including how and when these were defined)</td>
<td>2 (Verify file formats and other essential content characteristics)</td>
<td>Build relationships with content to encourage sustainable file or data practices</td>
</tr>
<tr>
<td>Content</td>
<td>2 (Document the file and other essential content characteristics, including how and when these were defined)</td>
<td>2 (Verify file formats and other essential content characteristics)</td>
<td>Build relationships with content to encourage sustainable file or data practices</td>
<td>2 (Build relationships with content to encourage sustainable file or data practices)</td>
</tr>
</tbody>
</table>

### 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.
| 1 - Content preservation | 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.
Sample with jet, track and secondary vertex properties for Hbb tagging ML studies
HiggsToBBNtuple_HiggsToBB_QCD_Runll_13TeV_MC
Duarte, Javier

Citation

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-anti-quark pair (Hbb) from quark or gluon jets originating from quantum chromodynamic (QCD) multiplet 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+H+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
Other Repository Integrations

- **figshare**
- **DSpace**
- **e-prints**

Diagram:
- Researcher
- Web browser
- HR system
- Local Research Data
- Figshare (Amazon)
- Journal
- DataCite (BL)
- Archive (Arkivum)
- CRIS (Elements)
- Repository (DSpace)

Connections:
1. Data files
2. Data files
3. Data Description
4. Mint DOI
5. Data DOI
6. Data DOI
7. Article
8. Data DOI
9. Article and Article DOI
10. Article and Article DOI
11. Article DOI
12. Dataset Description and Data DOI
13. Dataset Description and Data DOI
14. Data files
15. Data is safe
16. Data is safe
<table>
<thead>
<tr>
<th>Area</th>
<th>Short Requirement Name</th>
<th>FAIR Principle (Nature)</th>
<th>RDA Indicator</th>
<th>FAIR/FAIR Metric</th>
<th>Evidence Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digital Object Management</td>
<td>14. Data reuse</td>
<td></td>
<td>RDA-1-0-10 Data uses knowledge representation expressed in standardized format (important)</td>
<td>RDA-1-0-10 Metadata is represented using a formal knowledge representation language.</td>
<td>Metadata in the Arkivum solution is accessible in JSON or CSV format. Metadata can be searched and retrieved using a REST API. Metadata files are including JSON-LD and these metadata files can be associated to datasets, e.g., invenioRDM’s is in JSON-LD format and can be searched and exported via a REST API. InvenioRDM landing pages include record citation metadata in machine readable form (JSON).</td>
</tr>
<tr>
<td></td>
<td>11. Metadata uses a formal, accessible, shared, and broadly applicable language for knowledge representation</td>
<td></td>
<td>RDA-1-0-12 Data uses machine-understandable knowledge representation (important)</td>
<td>RDA-1-0-12 Metadata uses machine-understandable knowledge representation (important)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RDA-1-0-11M Metadata uses knowledge representation expressed in standardized format (important)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RDA-1-0-11M Metadata uses knowledge representation expressed in standardized format (important)</td>
<td>RDA-1-0-11M Metadata uses knowledge representation expressed in standardized format (important)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>I. Metadata use vocabularies that follow FAIR principles</td>
<td></td>
<td>RDA-2-0-1M Metadata uses FAIR-compliant vocabularies (important)</td>
<td>RDA-2-0-1M Metadata uses semantic resources.</td>
<td>Metadata files can be ingested into the Arkivum system in their native format (XML or JSON). These metadata files can include references to external schemes of vocabularies, e.g., schema.org.wikidata. InvenioRDM supports vocabularies: <a href="https://inveniosoftware.org-products/rdm/roadmap/">https://inveniosoftware.org-products/rdm/roadmap/</a></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RDA-2-0-1D Data uses FAIR-compliant vocabularies (useful)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>13. Metadata include qualified references to other metadata</td>
<td></td>
<td>RDA-3-0-1M Metadata includes references to other metadata (important)</td>
<td>RDA-3-0-1M Metadata includes references to other metadata (important)</td>
<td>The Arkivum solution supports Darwin Core relations 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: <a href="https://support.datacite.org/docs/datacite-type_for_arkivum-1">https://support.datacite.org/docs/datacite-type_for_arkivum-1</a></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RDA-3-0-2M Metadata includes references to other data (important)</td>
<td>RDA-3-0-2M Metadata includes references to other data (important)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RDA-3-0-2D Data includes qualified references to other data (useful)</td>
<td>RDA-3-0-2D Data includes qualified references to other data (useful)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RDA-3-0-3M Metadata includes qualified references to other metadata (important)</td>
<td>RDA-3-0-3M Metadata includes qualified references to other metadata (important)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RDA-3-0-3M Metadata includes qualified references to other data (useful)</td>
<td>RDA-3-0-3M Metadata includes qualified references to other data (useful)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R. Metadata are duly described with a plurality of accurate and relevant attributes</td>
<td></td>
<td>RDA-4-0-1M Plurality of accurate and relevant attributes are provided to allow reuse (essential)</td>
<td>RDA-4-0-1M Metadata specifies the content of the data.</td>
<td>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, Darwin Core, 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 vocabularies: <a href="https://inveniordn.dcc.ox.ac.uk/reference/metadata/resource-type-1">https://inveniordn.dcc.ox.ac.uk/reference/metadata/resource-type-1</a></td>
</tr>
<tr>
<td></td>
<td>13. Metadata meet domain relevant community standards</td>
<td></td>
<td>RDA-4-1-3M Metadata complies with a community standard (essential)</td>
<td>RDA-4-1-3M Metadata complies with a community standard (important)</td>
<td>Metadata files can be ingested into the Arkivum system in their native format (XML or JSON). In cases, 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.</td>
</tr>
</tbody>
</table>
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)
Open Standards, Open Specifications, Open Source: Deployment

- Kubernetes
- Terraform
- Ansible
- Rancher
- Grafana
- OpenStack
- Prometheus
Open Standards, Open Specifications, Open Source: Interaction

RClone

XRrootD

eduGAIN

research.object.org

Bagit & BDbags

s3fs

METS

XML
**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
## Preservation Formats and Access Formats

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

**Quick Links**

<table>
<thead>
<tr>
<th>Computer Aided Design</th>
<th>Digital Audio</th>
<th>Digital Moving Images</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digital Cinema</td>
<td>Digital Video</td>
<td>Digital Still Images</td>
</tr>
<tr>
<td>Digital Photographs</td>
<td>Scanned Text</td>
<td>Digital Posters</td>
</tr>
<tr>
<td>Geospatial Formats</td>
<td>Presentation Formats</td>
<td>Textual Data</td>
</tr>
<tr>
<td>Structured Data Formats</td>
<td>Email</td>
<td>Web Records</td>
</tr>
<tr>
<td>Calendars</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Preferred Formats**

<table>
<thead>
<tr>
<th>Preferred Formats</th>
<th>Format Versions</th>
<th>Format Specifications</th>
</tr>
</thead>
</table>

**Acceptable Formats**

<table>
<thead>
<tr>
<th>Acceptable Formats</th>
<th>Format Versions</th>
<th>Format Specifications</th>
</tr>
</thead>
</table>

| Word processors      |  |  |
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

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

As demand for data centers rises, energy efficiency improvements in 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 263 and 293 TWh by 2010, totaling 1.3 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 worldwide and compute instances had increased more than eightfold, whereas data center internet protocol (IP) traffic had increased by more than 10,600% (1). Data center storage capacity has also grown rapidly, increasing by an estimated factor of 23 over the same time period (1, 3). There has been a tendency among analysts to use such service demand trends to simply extrapolate earlier bottom-up energy values, leading to unfulfilled predictions of current and future global data center energy use (2–5). 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 2016) to arrive at an estimate of future energy use (e.g., total data center energy use in 2040).

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 process efficiency improvements and reductions in idle power (9). 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 (9). 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 efficiency (PUE), the total amount...
Green Data Centers, Renewable Energy
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
- Lower price
- Lower latency

Where is your traffic coming from?
- Ukraine
- United Arab Emirates
- United Kingdom
- United States
- Uruguay

Recommended regions

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

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

- northamerica-northeast1
  Montréal, Canada
  $ $ $
  - Grid carbon intensity: 27 gCO2eq/kWh
  3. 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

**Gross monthly carbon emissions**

- Data Start: Jan 2022
- Data End: May 2022
- Data Type: Kilograms of CO₂e
- Data Range: 0 to 2,000

**Gross carbon emissions by project in April 2022**

- Data Type: Kilograms of CO₂e
- Data Range: 0 to 2,000

**Gross carbon emissions by region in April 2022**

- Data Type: Kilograms of CO₂e
- Data Range: 0 to 2,000
### 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

<table>
<thead>
<tr>
<th></th>
<th>Ingest</th>
<th>Long-term storage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$2.2 per TB</td>
<td>$30 per TB-year</td>
</tr>
<tr>
<td></td>
<td>0.1 kgCO₂eq</td>
<td>0.7 kgCO₂eq per TB-year</td>
</tr>
</tbody>
</table>
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)
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

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

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

<table>
<thead>
<tr>
<th>Reporting</th>
<th>Read Only</th>
<th>User</th>
<th>Superuser</th>
<th>Admin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ingest Report</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
</tr>
<tr>
<td>Preservation Report</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
</tr>
<tr>
<td>Normalisation Report</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
</tr>
<tr>
<td>Hold Report</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
</tr>
<tr>
<td>Unhold Report</td>
<td></td>
<td>✅</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Processing Report</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Export Report</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Audit Trail</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deletion Report</td>
<td></td>
<td></td>
<td></td>
<td>✅</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Approval Workflows</th>
<th>Read Only</th>
<th>User</th>
<th>Superuser</th>
<th>Admin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Approve/reject ingests</td>
<td>✅</td>
<td>✅</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Approve/reject deletion requests</td>
<td>✅</td>
<td>✅</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Approve/reject Bulk Exports</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Authentication and Authorisation

https://vimeo.com/723625238/b5a3ee264a
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

Overall System Status

Location 1
- Dissemination data: 0TB
- Normalized Files: 0TB (16 files)
- Metadata: 0TB
- Audit Trail: 0.01TB
- Files: 47.32TB (833k files)

Data Usage
- 11.78%
- 47.33TB of 402.0TB

Current Processing
- 1A: Ingestion and integrity Checks
  - Completion: 100%
  - Process last completed: 2022-04-26 08:18:08
- PAE: Packaging and Encryption
  - Completion: 0%
  - Process last completed:
- MP: Metadata Processing
  - Completion: 0%
  - Process last completed:
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
<table>
<thead>
<tr>
<th>Datapool Name</th>
<th>Datapool Path</th>
<th>Preservation Enabled</th>
<th>Location Set</th>
<th>Metadata Namespaces</th>
<th>Webhook URL</th>
<th>Encrypted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arkivum</td>
<td>/arkivum</td>
<td>false</td>
<td>quickaccess</td>
<td>technical, <a href="http://www.arkivum.com/">http://www.arkivum.com/</a>...</td>
<td></td>
<td>true</td>
</tr>
<tr>
<td>Arkivum-Preserved</td>
<td>/arkivum-preserved</td>
<td>true</td>
<td>quickaccess</td>
<td>technical, <a href="http://www.arkivum.com/">http://www.arkivum.com/</a>...</td>
<td></td>
<td>true</td>
</tr>
<tr>
<td>Deep Archive</td>
<td>/deep_archive</td>
<td>false</td>
<td>deeparchive</td>
<td>technical, <a href="http://www.arkivum.com/">http://www.arkivum.com/</a>...</td>
<td></td>
<td>false</td>
</tr>
<tr>
<td>Default</td>
<td></td>
<td>false</td>
<td>quickaccess</td>
<td>technical, <a href="http://www.arkivum.com/">http://www.arkivum.com/</a>...</td>
<td></td>
<td>true</td>
</tr>
<tr>
<td>Frequent Access</td>
<td>/frequent_access</td>
<td>false</td>
<td>quickaccess</td>
<td>technical, <a href="http://www.arkivum.com/">http://www.arkivum.com/</a>...</td>
<td></td>
<td>false</td>
</tr>
<tr>
<td>ICDocs</td>
<td>/icdocs</td>
<td>false</td>
<td>quickaccess</td>
<td>technical, <a href="http://www.arkivum.com/">http://www.arkivum.com/</a>...</td>
<td></td>
<td>false</td>
</tr>
<tr>
<td>InvenioRDM</td>
<td>/invenio-rdm</td>
<td>false</td>
<td>quickaccess</td>
<td>dataciteTypes, dataciteDescriptions, te...</td>
<td><a href="http://invenio-rdm-1.arc">http://invenio-rdm-1.arc</a>...</td>
<td>true</td>
</tr>
<tr>
<td>MAGIC1</td>
<td>/magic1</td>
<td>false</td>
<td>quickaccess</td>
<td>technical, magic_... te...</td>
<td></td>
<td>true</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Datapool</th>
<th>Safeguarding</th>
<th>Preservation</th>
<th>InvenioRDM</th>
<th>Escrow</th>
<th>Read Only Access</th>
<th>User Access</th>
<th>Admin Access</th>
</tr>
</thead>
<tbody>
<tr>
<td>Datapool 01</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>User 1</td>
<td>User Group 1</td>
<td>Admin Group</td>
</tr>
<tr>
<td>Datapool 02</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>User 2</td>
<td>User Group 2</td>
<td>Admin Group</td>
</tr>
<tr>
<td>Datapool 03</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>User 2</td>
<td>User Group 3</td>
<td>Admin Group</td>
</tr>
</tbody>
</table>
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
Add new datapool

Name: Deep Archive
Path: /deep_archive

Encrypted: off
Preserved: off

Location set:
Deep Archive

Quota: 1 TB

Namespaces
Available namespaces:
- http://www.arkivum.com/xsd/atom
- http://www.arkivum.com/xsd/lsadg
- http://www.arkivum.com/xsd/esignature

Selected namespaces:
- http://www.arkivum.com/xsd/dublincore

* mandatory fields

Discard
Apply
## Metadata Namespaces

<table>
<thead>
<tr>
<th>Namespace</th>
<th>Prefix</th>
<th>Aggregation Types</th>
<th>Grouped</th>
<th>Editable</th>
<th>Mandatory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dublin Core</td>
<td>dc</td>
<td>All types</td>
<td>false</td>
<td>true</td>
<td>false</td>
</tr>
<tr>
<td>Usability</td>
<td>atom</td>
<td>All types</td>
<td>false</td>
<td>true</td>
<td>false</td>
</tr>
<tr>
<td>ISAD(o)</td>
<td>isodg</td>
<td>O</td>
<td>false</td>
<td>true</td>
<td>false</td>
</tr>
<tr>
<td>ESignature</td>
<td>eSignatures</td>
<td>F</td>
<td>true</td>
<td>false</td>
<td>false</td>
</tr>
<tr>
<td>Identifier</td>
<td>identifiers</td>
<td>All types</td>
<td>true</td>
<td>true</td>
<td>false</td>
</tr>
<tr>
<td>Technical</td>
<td>technical</td>
<td>F</td>
<td>false</td>
<td>false</td>
<td>false</td>
</tr>
<tr>
<td>DataCite Creators</td>
<td>dataciteCreators</td>
<td>C, F, O</td>
<td>true</td>
<td>true</td>
<td>false</td>
</tr>
<tr>
<td>DataCite Identifiers</td>
<td>dataciteIdentifiers</td>
<td>C, F, O</td>
<td>true</td>
<td>true</td>
<td>false</td>
</tr>
<tr>
<td>DataCite Types</td>
<td>dataciteTypes</td>
<td>C, O, F</td>
<td>false</td>
<td>true</td>
<td>false</td>
</tr>
<tr>
<td>DataCite Descriptions</td>
<td>dataciteDescriptions</td>
<td>C, O, F</td>
<td>true</td>
<td>true</td>
<td>false</td>
</tr>
<tr>
<td>DataCite</td>
<td>datacite</td>
<td>C, F, O</td>
<td>false</td>
<td>true</td>
<td>false</td>
</tr>
<tr>
<td>MAGIC Telescopes Stereo</td>
<td>magic_telescopes</td>
<td>F, C, O</td>
<td>false</td>
<td>true</td>
<td>false</td>
</tr>
</tbody>
</table>
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
<table>
<thead>
<tr>
<th>Location</th>
<th>Location Type</th>
<th>Cloud Provider</th>
<th>Bucket Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>cern-ingest</td>
<td>Ingest</td>
<td>GCP</td>
<td>arkivum-cern-ingest</td>
</tr>
<tr>
<td>cern-export</td>
<td>Export</td>
<td>GCP</td>
<td>arkivum-cern-export</td>
</tr>
<tr>
<td>XRootD Server</td>
<td>Export</td>
<td>S3 Compatible Storage</td>
<td>cern-export</td>
</tr>
<tr>
<td>Location 1</td>
<td>Content Archive</td>
<td>GCP</td>
<td>arkivum-cern-loca-1</td>
</tr>
<tr>
<td>Location 2</td>
<td>Content Archive</td>
<td>GCP</td>
<td>arkivum-cern-loca-2</td>
</tr>
</tbody>
</table>
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
Ingest Workflow using Buckets

Dataset

1. Upload
   - Data Files
   - Metadata Files (native format)
   - ark-manifest.json (metadata)

2. Ingest
   - Arkivum Solution

3. Cache

4. Long-term Storage A
   - Long-term Storage B
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
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

Bulk Export Request

Export Path (optional)

Location: CERN Export Bucket

Export Metadata Type: Content and Metadata

Export Format: BagIt

- Include URI in metadata export
- Include technical metadata in metadata export

Export Reports

<table>
<thead>
<tr>
<th>Export ID</th>
<th>ID</th>
<th>Export Status</th>
<th>Export Requested Time</th>
<th>Last Modified Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>62bd74a8f334e7b0d4d76b57c2</td>
<td>-</td>
<td>Success</td>
<td>2022-06-20 18:20:26</td>
<td>2022-06-20 18:24:54</td>
</tr>
<tr>
<td>62bd687334e7d04d7d4d0d0</td>
<td>C:\Calibrated____<em>_</em>_<em>_</em>_<em>_</em>_<em>_</em>_<em>_</em></td>
<td>Success</td>
<td>2022-06-20 18:22:25</td>
<td>2022-06-20 18:24:12</td>
</tr>
<tr>
<td>62bd57b3344e7b0d4d76b44f6</td>
<td>C:\Calibrated____<em>_</em>_<em>_</em>_<em>_</em>_<em>_</em>_<em>_</em></td>
<td>Success</td>
<td>2022-06-20 18:17:48</td>
<td>2022-06-20 18:22:51</td>
</tr>
<tr>
<td>62bd60e2334e7b0d4d76b5b5b</td>
<td>C:\Calibrated____<em>_</em>_<em>_</em>_<em>_</em>_<em>_</em>_<em>_</em></td>
<td>Success</td>
<td>2022-06-17 14:04:50</td>
<td>2022-06-17 14:09:09</td>
</tr>
<tr>
<td>62bd07c32dd4e7b0d4d76b503</td>
<td>C:\Calibrated____<em>_</em>_<em>_</em>_<em>_</em>_<em>_</em>_<em>_</em></td>
<td>Success</td>
<td>2022-06-17 01:00:02</td>
<td>2022-06-17 13:04:29</td>
</tr>
<tr>
<td>62bd2867b648e7b0d4d76b56f</td>
<td>C:\Calibrated____<em>_</em>_<em>_</em>_<em>_</em>_<em>_</em>_<em>_</em></td>
<td>Success</td>
<td>2022-06-15 07:29:38</td>
<td>2022-06-15 07:36:17</td>
</tr>
</tbody>
</table>

Cancel  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/ldoc-6948-1652866857/

Regards
Arkivum Support
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
NOT TO BE TAKEN AWAY
Metadata Example
(ark-manifest.json)

- **Identifier**
- **Type of entity** (Collection, Object, File)
- **Dataset Structure** (Parent - Child)
- **Location of Files** (e.g. with ingest)
Mapping to DublinCore

File containing native metadata

Json Path metadata extraction

Location of native metadata file

Metadata Example (ark-manifest.json)

```json
"entities": [
  {
    "id": "184881",
    "entityType": "C",
    "fileReferences": [
      "metadata/184881.json": { 
        "Key": "dc.title",
        "valuePathExpression": "$[0]['title']['title']"
      },
      "Key": "dc.subject",
      "valuePathExpression": "$[0]['subject']["term"]"
    }
  }
],
"id": "184881-metadata",
"entityType": "O",
"parentId": [ "184881"
],
"id": "184881-access-data",
"entityType": "O",
"parentId": [ "184881"
],
"id": "184881-original-data",
"entityType": "O",
"parentId": [ "184881"
],
"id": "metadata/184881.json",
"entityType": "M",
"url": "file:///metadata/184881.json",
"parentId": [ "184881-metadata"
],
"id": "access_copies/SCAN-9789037.pdf",
"entityType": "M",
"url": "file:///access_copies/SCAN-9789037.pdf",
"parentId": [ "184881-access-data"
]"}
Relationship to other entities

Where file should be in dataset hierarchy

Where to store the file

Type of relationship (DataCite relations)

Metadata Example (ark-manifest.json)
End Result
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

```json
"entities": [
  {
    "id": "12345",
    "entityType": "C",
    "fields": [
      {
        "key": "datacite.title",
        "value": "GluGluToHToTauTau dataset in reduced NanoAOD format for ed"
      },
      {
        "key": "dataciteDescribes.0.description",
        "value": "Dataset in reduced NanoAOD format for ed"
      },
      {
        "key": "dataciteCreators.0.familyName",
        "value": "Wunsch"
      },
      {
        "key": "dataciteCreators.0.givenName",
        "value": "Stefan"
      },
      {
        "key": "dataciteCreators.0.name",
        "value": "Wunsch, Stefan"
      },
      {
        "key": "dataciteCreators.0.nameType",
        "value": "personal"
      }
    ]
  }
]```
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
RetentionPolicy Schedules and Deletion

- Dataset deletion follows a request/approve workflow
- Deletion needs one approval
- 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
## Preservation Formats and Access Formats

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

#### Quick Links

<table>
<thead>
<tr>
<th>Computer Aided Design</th>
<th>Digital Audio</th>
<th>Digital Moving Images</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digital Cinema</td>
<td>Digital Video</td>
<td>Digital Still Images</td>
</tr>
<tr>
<td>Digital Photographs</td>
<td>Scanned Text</td>
<td>Digital Posters</td>
</tr>
<tr>
<td>Geospatial Formats</td>
<td>Presentation Formats</td>
<td>Textual Data</td>
</tr>
<tr>
<td>Structured Data Formats</td>
<td>Email</td>
<td>Web Records</td>
</tr>
<tr>
<td>Calendars</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Preferred Formats</th>
<th>Format Versions</th>
<th>Format Specifications</th>
</tr>
</thead>
</table>


<table>
<thead>
<tr>
<th>Acceptable Formats</th>
<th>Format Versions</th>
<th>Format Specifications</th>
</tr>
</thead>
</table>

|--------------------------------------|--------------------------------------------------|-----------------------------------------------|
File Format Normalisation
Long-term Archival Storage in Buckets

Dataset

- Data Files
- Metadata Files (native format)
- ark-manifest.json (metadata)

Upload

1. Upload

Arkivum Solution

2. Ingest

Cache

3. Cache

4. Long-term Storage A
   4. Long-term Storage B
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

Fixity Notifications

<table>
<thead>
<tr>
<th>Datapool</th>
<th>Last Fixity Time Check</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Artikum</td>
<td>2022-06-20 09:16:55</td>
<td>Clear</td>
</tr>
<tr>
<td>Artikum Preserved</td>
<td>2022-06-21 09:31</td>
<td>Clear</td>
</tr>
<tr>
<td>Deep Archive</td>
<td>2022-06-20 09:16:54</td>
<td>Clear</td>
</tr>
</tbody>
</table>
Datapools and Storage

https://vimeo.com/722930836/9659bb1784
Publication, Discovery, Analysis and Reuse of Research Data
Sample with jet, track and secondary vertex properties for Hbb tagging ML studies
HiggsToBBNtuple_HiggsToBB_QCD_RunII_13TeV_MC

Duarte, Javier

Citation

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) multiplet 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+Residual jet energy corrections are applied to the jets and pileup compensation is mitigated using the charged hadron subtraction (CHS) algorithm. Features of the AK8 jets with transverse momentum $p_T > 200$ GeV and pseudorapidity $|\eta| < 3.2$ are provided. Selected features of inclusive (both charged and neutral) PF candidates with $p_T > 0.95$ GeV associated to the AK8 jet are provided. Additional features of charged PF candidates (formed primarily by a charged particle track) with $p_T > 0.95$ GeV associated to the AK8 jet are also provided. Finally, additional features of
InvenioRDM Workflow

1. Prepare Dataset
2. Ingest
3. Call webhook
4. Retrieve DataCite Metadata
5. Create Record
6. Status Report
7. Review and Publish
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
# Self-Assessment using the CTS+FAIR Capability Maturity Model

<table>
<thead>
<tr>
<th>Area</th>
<th>Short Requirement Name</th>
<th>FAIR Principle (Nature)</th>
<th>RDA Indicator</th>
<th>FAIR+FAIR Metric</th>
<th>Evidence Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digital Object Management</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>13. Data discovery and identification</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discovery</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P2. Data are described with rich metadata (defined by R1 below)</td>
<td>RDA-F2.01M</td>
<td>Rich metadata is provided to allow discovery (essential)</td>
<td>F5-F2.01M</td>
<td>Metadata includes descriptive core elements (creator, title, data identifier, publisher, publication data, summary and keywords) to support data findability.</td>
<td>Metadata is supported in the Arkivum solution includes publications, datasets, 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 Dublin Core and then made searchable.</td>
</tr>
<tr>
<td>Discovery</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P4. Metadata are registered or indexed in a searchable resource</td>
<td>RDA-F4.01M</td>
<td>Metadata is offered in such a way that it can be harvested and indexed (essential)</td>
<td>F5-F4.01M</td>
<td>Metadata is offered in such a way that it can be retrieved by machines.</td>
<td>DataCite metadata can be exported to InvenioRDM.</td>
</tr>
<tr>
<td>Identification</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P1. Metadata are assigned a globally unique and persistent identifier</td>
<td>RDA-F1.01M</td>
<td>Metadata is identified by a persistent identifier (essential)</td>
<td>F5-F1.02D</td>
<td>Data is assigned a persistent identifier.</td>
<td>All entities in the Arkivum solution (PCDM collections, objects, files) are assigned globally unique UUIDs.</td>
</tr>
<tr>
<td></td>
<td>RDA-F1.01D</td>
<td>Data is identified by a persistent identifier (essential)</td>
<td>F5-F1.02D</td>
<td></td>
<td>Users can provide one or more of their own identifiers as metadata, e.g., PIDs such as DOI or Handles. The Arkivum metadata schema supports identifier type and identifier value.</td>
</tr>
<tr>
<td></td>
<td>RDA-F1.02M</td>
<td>Metadata is identified by a globally unique identifier (essential)</td>
<td>F5-F1.02D</td>
<td></td>
<td>URLs are possible to entities in the Arkivum solution (authentication is required for access).</td>
</tr>
<tr>
<td></td>
<td>RDA-F1.02D</td>
<td>Metadata is identified by a globally unique identifier (essential)</td>
<td>F5-F1.02D</td>
<td></td>
<td>Records in InvenioRDM can have one or more PIDs, including those registered with external PID systems or as Alternative Identifiers. Identifier schemes include: ARK, mivio, Handle, DOI, EAAV, ESGR, Handle, ESR, ESRN, IST, IST, LSOR, LSID, Public ID, PID, CPC, URL, URN, DOI.</td>
</tr>
<tr>
<td>Identification</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P3. Metadata clearly and explicitly include the identifier of the data it describes</td>
<td>RDA-F3.01M</td>
<td>Metadata includes the identifier for the data it describes</td>
<td>F5-F3.01M</td>
<td>Metadata includes the identifier of the data it describes.</td>
<td>Metadata in InvenioRDM includes identifiers (see above). Likewise, all records in the Arkivum solution have identifiers. Records in InvenioRDM can include files. The InvenioRDM roadmap includes support for referencing data in external systems.</td>
</tr>
</tbody>
</table>
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
QUESTIONS?

matthew.addis@arkivum.com
orcid.org/0000-0002-3837-2526
www.arkivum.com