

Roadmap : 2025-2030 EBI Storage strategy

Strategic Plan: Retiring POSIX-based Storage

Vision

1. Strategic Rationale

2. Strategic Phases

Phase 1: Public POSIX Migration (Pilot with Isilon Public)

Phase 2: FIRE Project Migration to Weka

Phase 3: Build the ITS Modern Data Portal

Phase 4: Governance, Monitoring, and Lifecycle Automation

3. Technical Architecture

4. Risks and Mitigations

5. Governance & Milestones

6. Narrative Summary

Annex

Extra ideas to think about

Raw plan

Based on:  [Strategic Plan: Retiring POSIX-based Storage at EMBL-EBI](#)

Strategic Plan: Retiring POSIX-based Storage

Vision

To modernise EMBL-EBI's storage infrastructure by phasing out legacy POSIX-based systems (e.g. Isilon), consolidating data access through scalable, object-based storage (Weka), while preserving performance, user experience, and archival integrity.

1. Strategic Rationale

- **Scalability:** Data is growing at >1PB/month.
- **Vendor Agnosticism:** Avoid lock-in through object storage that decouples metadata and access paths.
- **Cost Efficiency:** Leverage Weka's HSM and Point Archival Gateway to move cold data to tape without rehydration costs.
- **Resilience and Transparency:** Replace brittle mount-based access with stable S3 semantics, preserving existing URLs and access paths through virtualisation layers.

2. Strategic Phases

Phase 1: Public POSIX Migration (Pilot with Isilon Public)

-  **Scan & Virtualise:** Weka scans Isilon public namespace.

- 🔄 **Remount:** All systems that are currently using Isilon Public will remount with the Weka native POSIX client.
- ✅ **Transparent Transition:** S3 endpoint becomes available for the same data, existing access paths are preserved.
- ✅ **Isilon Retirement:** Isilon phased out, with backend content now served by Weka.

Outcome: Users retain legacy path access; ITS gains object interface and cloud-friendly storage.

Phase 2: FIRE Project Migration to Weka

- 📦 **Per Project Scan:** Weka scans each FIRE bucket/project namespace.
- 🧠 **FIRE Logic Externalised:** Metadata and archival logic (OID/replica states) extracted or emulated within the Weka namespace context.
- 🔄 **Mount Projects with Weka native client:** Access projects via Weka's POSIX client, maintaining a consistent user experience with the rest of the storage systems.
- 🔒 **Ensure DR Parity:** All data must still be replicated (Weka + PoINT to tape).
- ❌ **Retire FIRE stack gradually:** POINT (tape) integration retained behind Weka S3 interface; FIRE's replication will be deprecated.

Outcome: FIRE metadata and archival workflows persist, but physical infrastructure becomes software-defined on Weka and standardised with the Public isilon replacement approach.



Phase 3: Build the ITS Modern Data Portal

- 🌐 **Portal as a Service:** Globus-powered frontend backed by Weka's S3.
- 🔑 **Federated Login:** Users can authenticate using Globus and SAML, eliminating the need for multiple local accounts..
- 🧩 **Code Reuse:** Provide pre-built upload components to ENA/EGA and others.
- 🔄 **Unify Ingress Pathways:** Promote the Portal as the standard EBI-wide data ingestion point.

Outcome: Scalable, user-friendly, authenticated upload platform. Reduces redundant systems and improves security. Specific value for smaller teams and projects, as Unified Submission Interface intended.

Phase 4: Governance, Monitoring, and Lifecycle Automation

- 📄 **Metadata-driven Lifecycle:** Use metadata or access logs to drive HSM tiering in Weka.

-  **Unified Observability:** Integrate Weka, Globus, and archival system metrics (PoINT tape) into a single dashboard.
-  **Automation Pipelines:** Auto-migrate unused datasets to cold storage based on policies.

3. Technical Architecture

Layer	Technology	Role
Object Storage	WekaIO	Primary, fast, scalable store
POSIX Interface	Weka Kernel Module	Legacy-compatible access
S3 API Layer	Weka S3 Gateway	External and internal object access
Cold Archive	PoINT + Tape (OTA)	Long-term disaster recovery
Portal Frontend	Globus + SAML	Federated upload & auth
Monitoring	Fluentd + Grafana	Observability, capacity planning

4. Risks and Mitigations

Risk	Mitigation
Weka vendor lock-in	Preserve open metadata formats; use standard S3, and isolate archival logic
User disruption	Maintain file paths via consistent mount points; stage remounts
Cold data inaccessibility	Keep tape index searchable via Weka and PAG
IAM complexity	Use Globus SSO and role-based access control from Day 1

5. Governance & Milestones

Quarter	Milestone
Q3 2025	Weka public namespace live, Isilon read-only
Q4 2025	FIRE buckets accessible via Weka kernel client
Q1 2026	Modern Data Portal MVP launched (Globus + Weka)
Q2 2026	First FIRE project retired; Portal used by 3+ teams
Q3 2026	PAG accessible via Weka S3; FIRE replication to PAG decommissioning begins

6. Narrative Summary

EMBL-EBI's storage infrastructure has evolved from legacy POSIX filesystems to one of the largest scientific archives in the world. Now, it faces the next stage in its evolution: a shift to scalable, software-defined object storage. This strategy transitions EMBL-EBI's data platform into a future-proof, user-friendly, and cloud-native system while retaining legacy compatibility, maximising open access, and reducing costs through automation and modern lifecycle management.

Annex

Extra ideas to think about

- Globus can use an S3 backend instead of POSIX, would that deliver better or worse performance than the Weka native client?
- Aspera HSTS does not support S3 natively, but it does through goofys. We shall expect the Weka native client to perform better than goofys; *cache* is the key element.

Raw plan

Steps to get into an object storage state:

1. Isilon public mount points to Weka Public namespace
 - a. Weka scans existing Isilon public content, and all systems mounting public isilon remount on the agreed day to the Weka native kernel module.
 - b. Weka S3 interface is gained for all that data while keeping existing isilon attributes.
 - c. Isilon can be retired behind the scenes of the Weka system, transparently to the end user.
 - d. Weka can move data to cheaper storage, as it has HSM capabilities. **Cost reduction**
2. FIRE projects migrated to WEKA
 - a. FIRE team implement metadata writing to HGST with Weka agreement to read it on similar scan (like they did with isilon)
 - b. For each FIRE project
 - i. Weka scans the buckets of the project.
 - ii. Weka native kernel module is used to mount project with same approach as with the Weka Public namespace
 - c. Retire the FIRE project with an off-the-shelf solution like Weka. No vendor lock-in, as Weka keeps metadata on disc, and another system could do what Weka did with Isilon: scan and replace.
3. ITS branded Modern Data Portal, from Globus, where ITS provides a opinionated approach to upload data to EBI, leveraging directly Weka S3 system
 - a. Allows all small projects to leverage this system without having to create a new one.
 - b. Is provided as an internal codeset for other teams, like ENA and EGA, to use and collaborate with.
 - c. Leverages Globus SAML authentication, reducing internal IAM accounts.

- d. Over time it should be the system used to upload data into EBI storages.
- e. Unified submission interface was a first iteration; this time the IAM systems would be ready.

4.