Guest author(s)

5 June 2024

The EMBL-EBI File Replication Archive hits 100 petabytes

A behind-the-scenes look at EMBL-EBI's in-house archive that supports major open data resources and ingests record amounts of data

Credit: Photo by Jeff Dowling. Editing by Karen Arnott.

Behind the scenes, the EMBL-EBI File Replication (FIRE) archive has, for over 15 years, been enabling thousands of scientists to share and access data openly to advance their work. In turn, this has supported the development of solutions for global challenges in infectious disease, food security and biodiversity, with truly global impact. An independent study from 2021 found that EMBL-EBI data and services contribute to the wider realisation of future research impacts worth £2.2 billion annually – and since then, usage of our data resources by scientists across the world has continued to grow.

What is FIRE?

FIRE is an internally developed, software-defined, geo-dispersed storage system, in which some of the most important data held at EMBl-EBI is stored (you can read more about it here).

FIRE provides a home for some of EMBL-EBI's largest and most popular data resources. These include the European Nucleotide Archive (ENA), the European Genome-phenome Archive (EGA), the PRoteomics IDEntifications Database (PRIDE), and the Bioimage Archive. These critical resources must remain highly available, performant and secure, and FIRE provides these things via mechanisms developed by the teams of our IT & Technical Services department.

FIRE sits at the heart of EMBL-EBI's vast library of scientific data, and this latest achievement highlights the growing importance of EMBL-EBI data resources to the research community. What follows is a potted history of FIRE through the years, along with some of the challenges it has faced and the solutions implemented. If you want to know how we got here, read on...

The early years

FIRE was 'born' on February 5 2009, initially providing a home for the ENA and the 1000 Genomes project, with EGA joining in July 2010. Three and half years later, in July 2012, FIRE had ingested its first whole petabyte of data, and as demand for the services grew, this increased dramatically to 10PB by April 2016. During these early years, FIRE was managed by our Infrastructure team, with just a single member of staff dedicated to developing and maintaining it. In 2017, with the help of increased funding, a project team was formed within the ITS Operations team. The extra resource allowed us to completely revamp the existing software stack, from its original Python-based implementation to Java, deployed using Spring Boot on Kubernetes. FIRE continued its rapid growth trajectory, reaching a cumulative size of 20PB in September 2018.

In 2019, the growing importance of FIRE at EMBL-EBI led to the formation of the FIRE Users' Group, drawing input and collaboration from across the institute. "At the first Users' Group meeting we shared our plans to improve FIRE by introducing an S3-

compatible read-only FUSE implementation," said Marc Riera, FIRE Team Coordinator, significantly reducing complexity within the software, and making FIRE easier to use, with improved error handling. Marc continues, "Shortly after the BioImage Archive joined in 2019, we added an HTTP-based blob uploader to FIRE, which allowed the services to push data to us easily and quickly via HTTP – a big step forward."

Keeping up with demand

By the end of 2019, unrelenting growth combined with the improvements to data transfer speeds, were taking their toll on the network. "The increased traffic as a result of the S3 reads was starting to cause the network to struggle, so our colleagues in the Networking team had to double the capacity between our on-site and off-site data centres," recalls Marc. December 2019 saw FIRE reach a new record of 70TB ingressed in a single day via the new HTTP API, smashing the previous record of 47TB. Both ENA and EGA would switch to using the new write system in 2020, resulting in a new one-day ingress record of 115TB in May.

> "The achievement of ingesting 100 petabytes into FIRE is a remarkable milestone and a testament to the innovative work of the IT and science teams at EMBL-EBI. FIRE's ability to seamlessly scale and support the ever-growing needs of the scientific community underpins EMBL-EBI's mission to advance life sciences research."
>
> Andy Cafferkey, Head of IT & Technical Services

FIRE keeps a replica of everything on tape, for disaster recovery purposes. By August 2020, the constant and overwhelming demand for FIRE required the tape backup system to quickly grow way beyond its original remit, with multiple different tape technologies in use and an internally developed object storage stack that could no longer cope. This system was completely replaced in August 2020 with an S3-compatible Point Archive

Gateway system, which provides a simplified and standardised solution, that is both more performant and cost-effective.

The future

Over its lifetime, FIRE has inevitably undergone numerous hardware and storage system changes and upgrades, and it does this – crucially – in a way that is transparent to users. “Data from 2008 may have been stored initially in an NFS storage from one vendor, then moved onto NFS from a second vendor, then to an early object store, and finally to the current modern object store technology,” explained Marc. Throughout that time, and regardless of those changes, the archives have appeared identically to the users accessing them – “the same file, the same file path, but with a constantly evolving underlying storage technology,” said Marc. This hiding of the hardware lifecycle has been central to FIRE’s success as an attractive solution for the long-term archiving of important data. Continued transformation in EMBL-EBI’s data infrastructure is being supported by UK Research and Innovation’s (UKRI) Infrastructure Fund.

FIRE continues to go from strength to strength, welcoming PRIDE onboard in September 2021 and then BioStudies in February 2022. The data onslaught hasn’t relented since, consistently surpassing all predictions – especially in the wake of the pandemic, which has sparked new interest in life sciences research across the globe. “The plan is to replace the current HTTP API with S3, which is the final piece of the standardisation puzzle...for now,” concluded Marc.

The importance of FIRE

The long-term data preservation, service continuity and robustness that FIRE supports are among the reasons that EMBL-EBI data resources are trusted – by scientists who want their data to remain Findable, Accessible, Interoperable and Reusable (FAIR), and by funders who need trustworthy repositories to support open science and promote reproducible research practice by their awardees.

As bodies such as the Global Biodata Coalition and the European Commission develop principles and models for ensuring the long-term sustainability of critical data resources, it becomes even more important for data resources to communicate how systems like FIRE contribute to the preservation, curation, and security of valuable research data.

> “EMBL-EBI is now among the world’s leading providers of open biological data resources at scale. Researchers across the life sciences trust and rely on EMBL-

EBI data resources, and FIRE is a big contributor to earning that trust – ensuring we keep their data secure in the long-term and provide our users with reliable and performant data services."

Jo McEntyre, EMBL-EBI Deputy Director

Related links

EMBL-EBI IT and Technical Services

Technical careers at EMBL-EBI

Technical jobs at EMBL-EBI

IT and Technical Services blog

Tags: archive, bioinformatics, embl-ebi, information technology, technology,

Share this

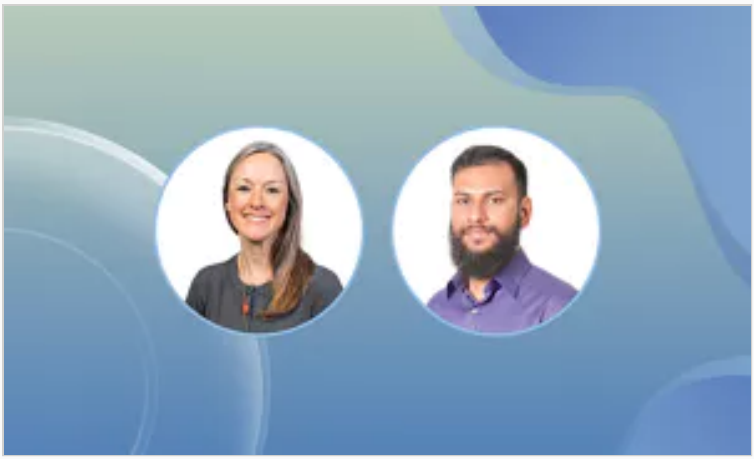
More Technology and innovation

GA4GH announces refget Sequence Collections

A federated future to support genomic medicine

27 March 2025

7 March 2025



Understanding rare diseases with digital twins

5 March 2025

Polygenic Score (PGS) Catalog increases diversity and usability of genetic data

26 September 2024

Contact Industry team

ABOUT

Contact us

Events

Jobs

News

People and groups

Intranet for staff

EMBL-EBI, Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK. Tel: +44 (0)1223 49 44 44 Full contact details