EMBL-EBI Purchasing and Strategic Direction 2022 - 2023

EMBL-EBI

Tim Dyce

Head of Systems Infrastructure

tim.dyce@ebi.ac.uk

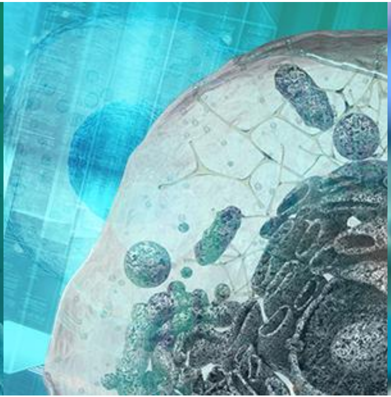
EMBL-EBI

What is EMBL-EBI?

- World leading source of public biomolecular data
- Our vision is to benefit humankind by advancing scientific discovery and impact through bioinformatics.
- Part of the European Molecular Biology Laboratory (EMBL), Europe’s flagship laboratory for the life sciences.

Our mission

Deliver data resources



Perform excellent research

Train the next generation of scientists

Engage with industry

Coordinate bioinformatics in Europe

Sources of funding

- EMBL-EBI is primarily funded by EMBL's member states.
- Other major funders:
  - European Commission
  - UK Research and Innovation
  - MRC
  - National Institutes of Health
  - Wellcome
  - Industry Programme

EMBL-EBI Infrastructure

As leading source of public biomolecular data EMBL-EBI has a strong focus on data storage and services

The largest infrastructures at EMBL-EBI are:

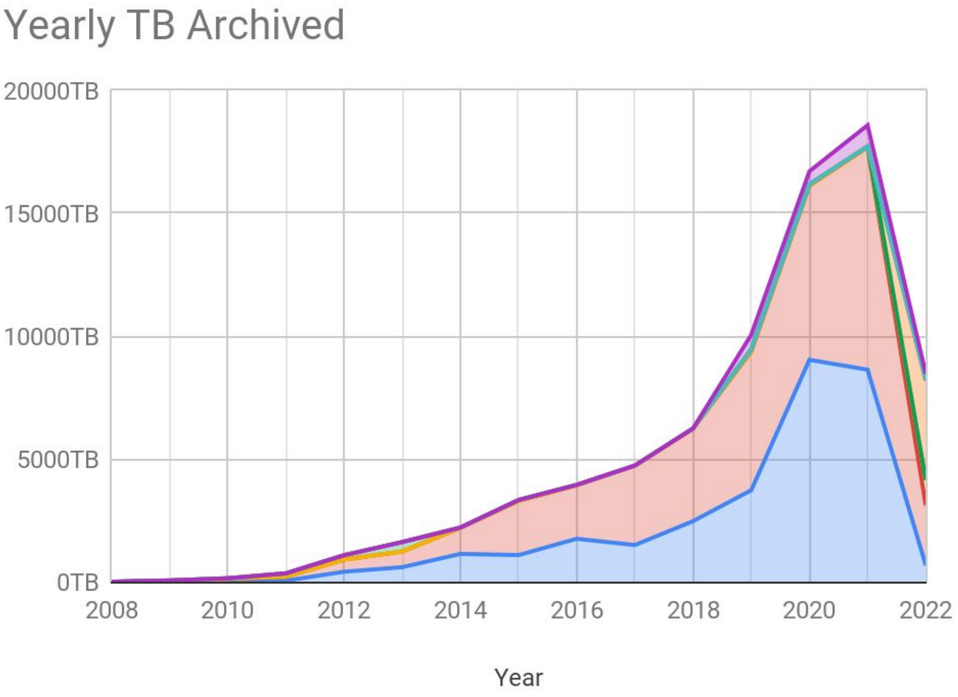
- 3 Geo. object storage (~ 90PB usable)
- NAS/POSIX storage (~ 30PB usable)
- Tape archive (~90 PB usable)
- A 300 node HPC cluster

The volume of data archived by EMBL-EBI increases by around 30% annually

Operating at this scale EMBL-EBI tends to mainly deal with providers with multiple deployments at the scales described

Storage Growth

Year	Daily growth	Yearly Growth	Archive Size	Data Size
2021	62 TB	22 PB	67 PB	201 PB
2022	94 TB	33 PB	100 PB	301 PB
2023	140 TB	50 PB	151 PB	452 PB
2024	211 TB	75 PB	226 PB	677 PB
2025	316 TB	113 PB	338 PB	1015 PB



Strategic Direction

2021 - 2022: Consolidation

- Rationalisation of our large NAS/POSIX estate
- Completing migration to new a DC and HPC cluster
- Evaluation of secure computation environments
- Evaluation of future storage technologies

2023: Optimization

- Starting transition to direct use of object storage
- Further leveraging cloud
- Cloud accessibility to large archives
- Offline storage for inactive data
- Review of remote access technologies

2024: Transformation

- Data management tooling and cataloging
- Next generation network microsegmentation to better support controlled access data

Future Needs

2023

- Large scale object storage
  - 40PB+
- Tape library hardware and tape media
  - 30PB+
- S3 API compliant tape management software and tooling
  - 100PB+

2024

- Similar to 2023

2025

- A replacement HPC cluster environment

Scale and Volume of Investment

YEARS	Forecast (million EUR)
2023	13.1
2024	12.9
2025	14.9
2026	15.1
2027	17.3

- EMBL-EBI procurements follow defined internal EMBL financial rules. Requirements >EUR 12.5k follow a competitive procedure, where possible
- Typically, complex solutions are tendered using pre defined weighted award criteria (price, quality and sustainability) to pre selected vendors.
- EMBL-EBI also procures hardware through established commercial frameworks
- EMBL-EBI proactively engages with the wider market identifying opportunities with new technology and suppliers
- An external EMBL procurement webpage is currently in development, including publishing relevant information for suppliers to express their interest in becoming an approved supplier

Lessons Learned

Many storage, data management and analytics products tailored to life sciences data offer attractive features, however:

- Most advanced features do not scale well beyond 20PB and several billion files
- Storage and tools at this scale often need to be simple to be sustainable

Direct access to engineers and developers makes operation of large scale infrastructures simpler for both EMBL-EBI and the vendor