Science and Technology Facilities Council

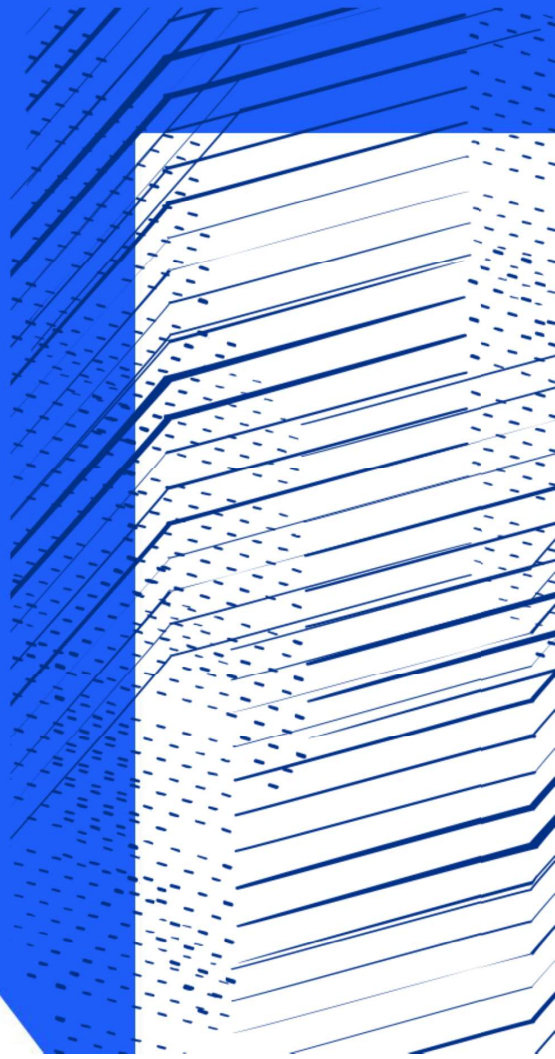
‘5 mins on Data Management Challenges’... & Opportunities

Tom Burnley CCP-EM/STFC

19 July 2024

UK Community Consultation on Cryo-EM

CCP-EM



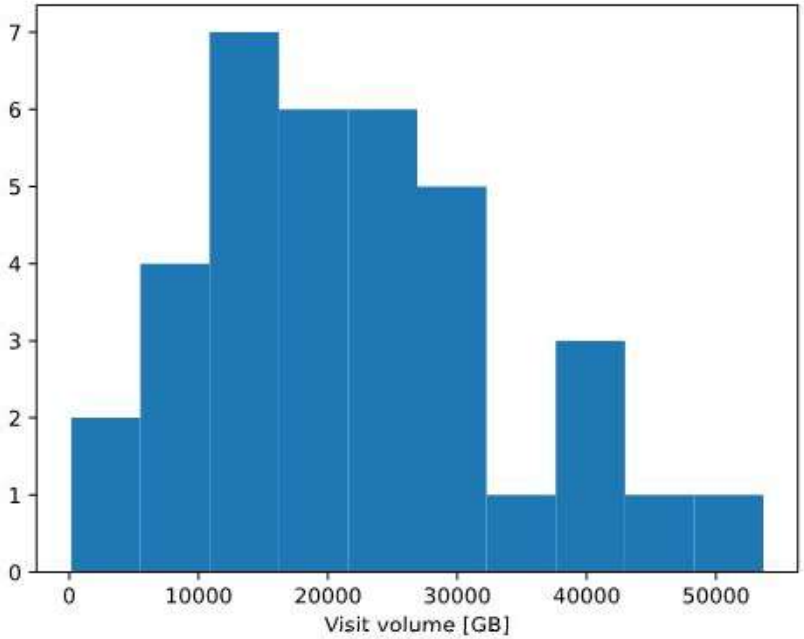
How much data in 30 days?

- SPA | 18 TB* per visit | 165 visits | 2.9 PB total
  ~9TB raw data
- Tomo | 3.6 TB per visit | 64 visits | 0.28 PB total

- Current yearly estimate ~1.5 PB per Krios (4)
- Total 2023 archive 9.5 PB

- ~25 other high end microscopes in UK
- Need solutions for all

- 400+ microscopes world wide
- SKA projected to produce 1PB / day



SPA: Krios I – 22 TB (36 visits)
(-16/7/24)

Science and Technology Facilities Council

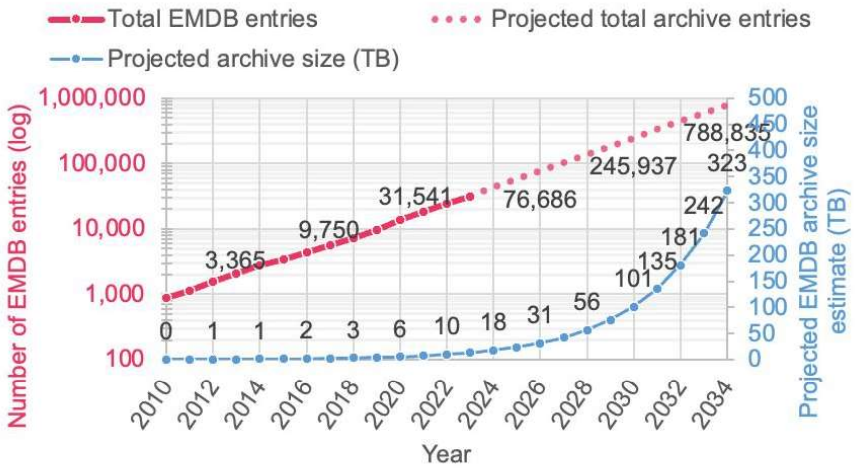
diamond

Dan Hatton DLS

How much in EMDB & EMPIAR?


Kyle Morris EBI


Matthew Hartley EBI

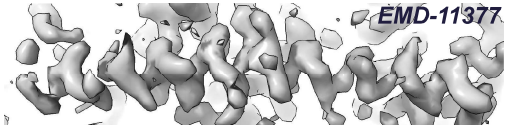


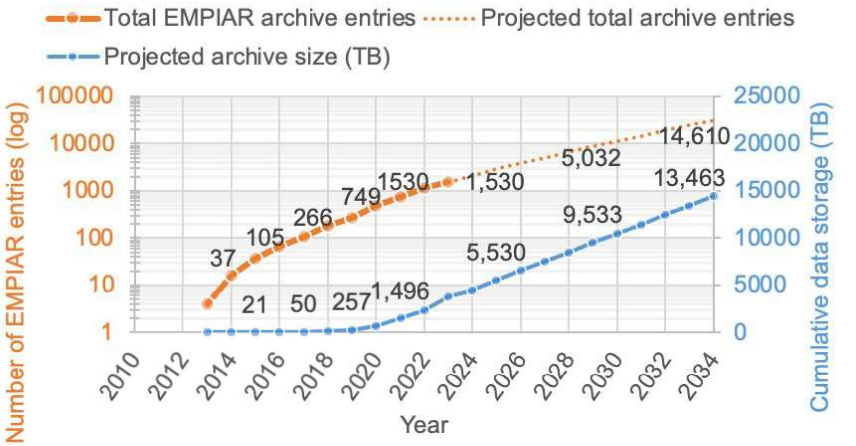
- ***Exponential EMDB archive growth***

	2024	2034
Total entries	37,053	788,835
Archive size (TB)	18	323

Weekly throughput!
~2,916 entries per week

EMDB
Electron Microscopy Data Bank



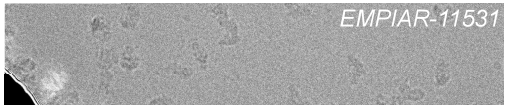


- ***Sustained EMPIAR archive growth***

	2024	2034
Total entries	2,040	18,939
Archive size (PB)	4.3	13.5

Weekly throughput!

EMPIAR



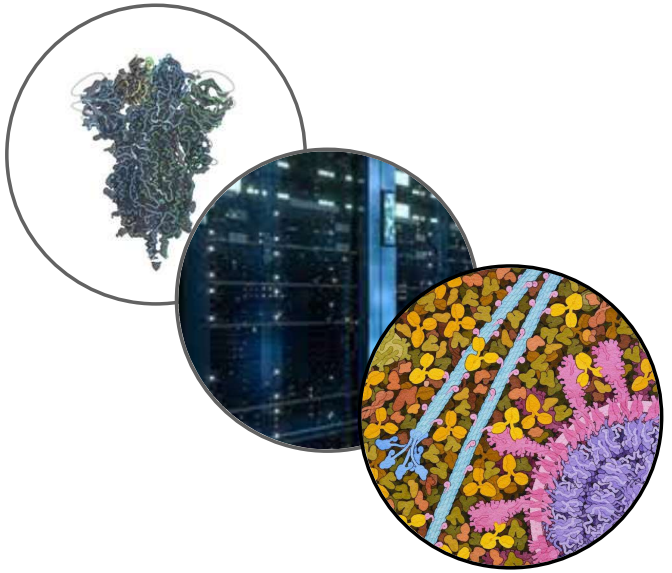
EMBL-EBI

Data complexity - EMDB & EMPIAR

- To realise the full value of structural biology data in the future, data complexity needs to be captured in the databases = we need to resource talent to build the systems to do this, not just hard drives.

- Data has context
  - Cellular context
  - Time resolution
  - Conformational landscape/HRAs

- Specimens could be connected to many other data types
  - Cross-linking
  - Sequencing
  - Proteomics
  - …



UKRI Science and Technology Facilities Council

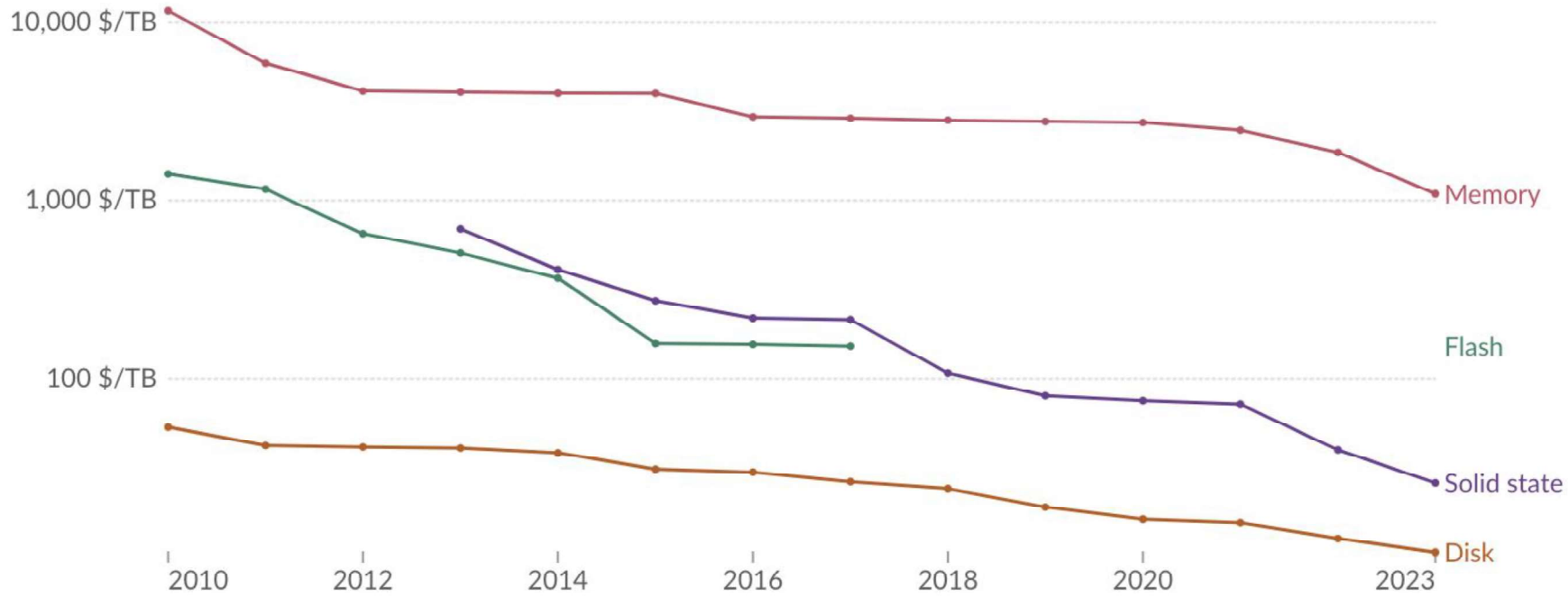
EMBL-EBI

Matthew Hartley EBI

Kyle Morris EBI

Data storage costs

- Reducing but likely not sufficient for increase in cryoEM data acquisition



Price of computer memory and storage

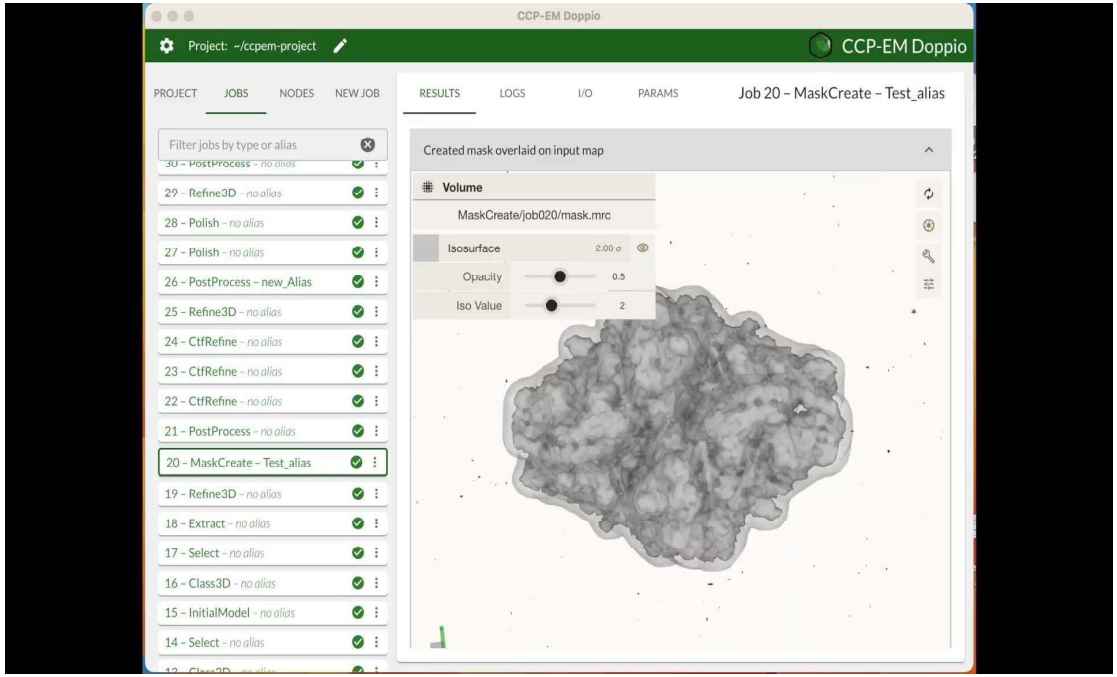
Science and Technology Facilities Council

https://ourworldindata.org/grapher/historical-cost-of-computer-memory-and-storage?time=2010..latest

Metadata gap

- FAIR ‘findable’, accessible, interoperable and re-useable
- Metadata required for efficient re-use of stored data, AI
- *Efficient metadata deposition requires automation*
- *CCP-EM Doppio, Scipion, others have metadata tools but need to finish links with facilities and repositories*

- **EMPIAR**
  - Entries with micrographs: **1679**
  - Entries with no metadata: **1578 (94%)**
  - ~15% can be reprocessed automatically
  - ~40% reprocessed with manual intervention



CCP-EM Doppio

- *Raw data ~GB-TBs | Metadata ~KB-MBs*
- *Cost not in storage but investment to produce metadata gathering, annotation and deposition pipelines*

Matt Iadanza CCP-EM

Science and Technology Facilities Council

Challenges & opportunities

- CryoEM in very good position (resources & culture) compared with other domains but we could, and will need to, do better…

- Is it possible to store ‘everything’ anymore?
- Should we store all datasets or limit to ‘productive’ datasets?
- What data should be stored?
- Audit storage (optimise formats, compression, heterogeneous storage)?

- Automate collation of metadata
- Enrich archives by linking to others
- Data stewardship adds reach and value beyond original study and community
  - ***<0.1% PDB users are experimental structural biologists***
  - ***What data & metadata will drive future breakthroughs?***