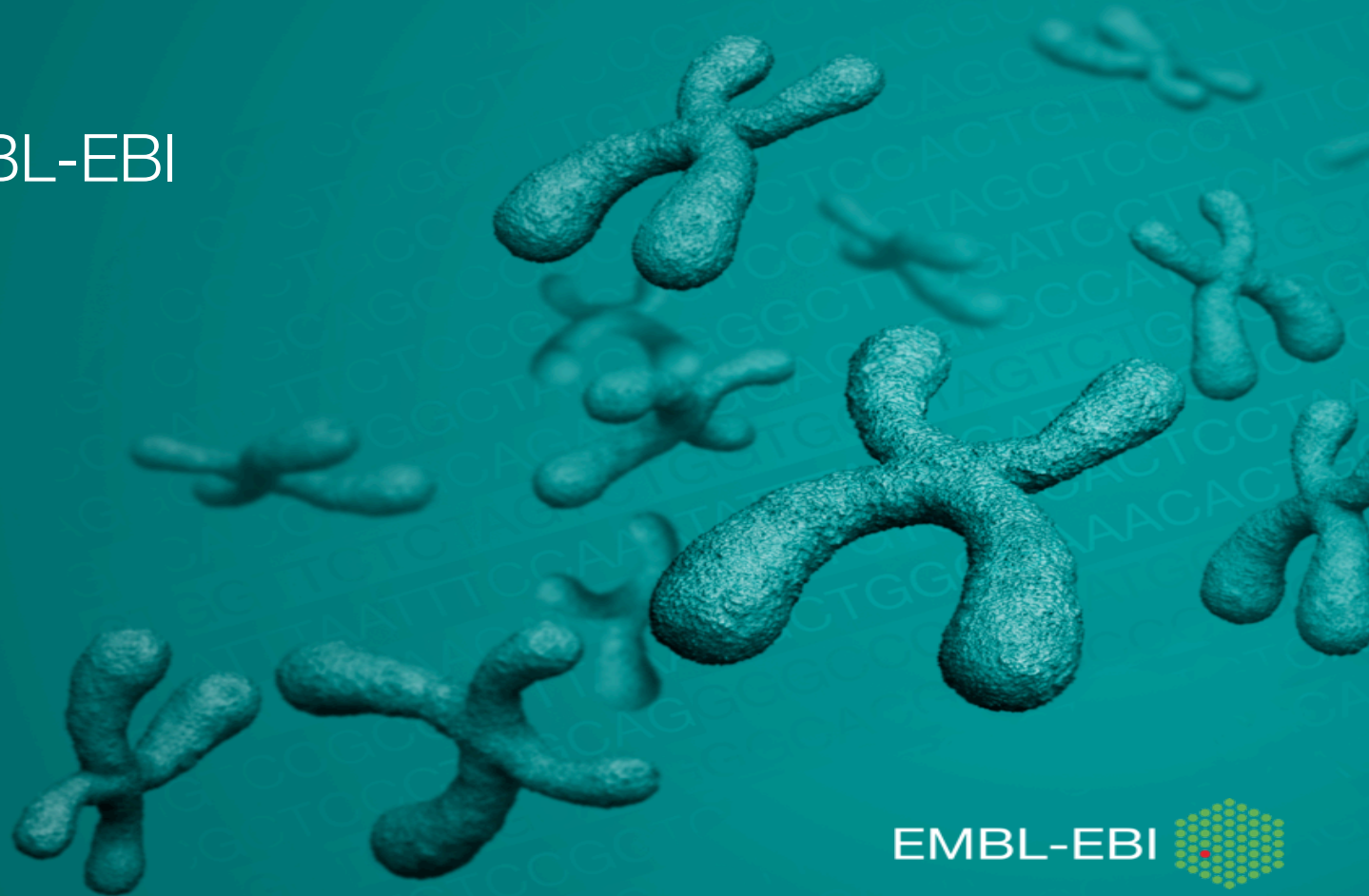


# Data Strategies for Research Infrastructures: Financial Requirements

Steven Newhouse

Head of Technical Services, EMBL-EBI



## A Recipe for Happiness (in Life and Research Infrastructures)

Annual income twenty pounds, annual expenditure nineteen [pounds] nineteen [shillings] and six [pence], result happiness.

Annual income twenty pounds, annual expenditure twenty pounds ought and six, result misery.

- Charles Dickens, David Copperfield

# My Recent Experiences

- European Grid Infrastructure
  - Transitioning from project to income based services
  - What are the core services needed to sustain them?
  - How much do they really cost to develop, operate & support?
- European Bioinformatics Institute
  - Commercial and collaborative use of the Embassy Cloud
  - Cost of the Technical Services Cluster

# Expenditure

- Capital Costs
  - Buildings
  - Hardware
  - Software
- Operating Costs
  - People
  - Software
  - Support
  - Physical & Electronic Infrastructure

# EMBL-EBI Embassy Cloud

- Basic Embassy (annual): £39K
  - 10GHz CPU, 64GB RAM, 1TB HDD
- Increasing resources
  - £2700/TB of storage
  - £1800/core of CPU or £720/GHz of CPU
  - £70/GB RAM
- Allowance for:
  - Staff management, hardware, operating costs, software, ...
- No allowance for:
  - Consultancy (beyond start up hand holding), additional system administration, ...

# EGI Core Services

- Transitioning from a project to a payment structure
  - What are the services?
  - What do they cost?
  - How important are they and to who (users, funders, operations, ...)?
- Result of the analysis
  - Divide services into: Essential, Useful, Nice to have, Not Needed
  - EGI.eu community pays to operate and maintain Essential services
  - Project funding to develop all services and operate all
    - New enhancements and non-essential services dependent on funding success

# EBI Technical Services (ongoing)

- Establishing and defining a service portfolio
  - See next slide: Split between external and internal services
- Understand the cost metrics behind each service
  - Usage: CPU hours, PB storage, GB network traffic, ...
  - Metrics: # people (network infrastructure), # machines (desktop)
- Understand the costs
  - Use to modify behaviour? [Fast Storage, Network Storage, Tape Storage]
  - Use to inform management? [Align costs with priorities]

# Service Portfolio

- Cloud: Embassy Cloud, Self-Managed VM, TSC Managed VM
- Database: SQL Database, No-SQL Database
- Desktop: Desktop Application Support, Desktop Hardware
- Web Production: Fully managed/Supported/Un-Supported Web VM
- Web Platforms: EBI Search, Job Dispatcher
- Web Infrastructure: Source Code Management System, Document Management, Team Communication and Management, Issue Tracker, Help desk
- Web Development: Project Website, User Experience Consultancy, Web Development
- Storage: Online Storage, EBI Archives, Sequence Retrieval Archive
- Compute: Batch Computing Cluster
- Network: External Network Services, Core Network Services



# Income from Grants

- CapEx: Many funding bodies not yet shifted from CapEx to OpEx
  - BBSRC currently generously supports EBI CapEx
  - EBI currently supports EBI OpEx for technical staff (limited people)
    - Some growing support through grants, but requires change in working practices
- OpEx: Covering people
  - Systems Administration costs (small, predictable but not trivial)
  - Consultancy & Support (potentially large & unpredictable)
- EC: If you can provide audited costs (people & OpEx)
  - Transnational Access: Costs of operating access to precious resource
  - Virtual Access: Cost of operating a resource for everyone

# Income from Fees

- Minimise the amount you need to collect
  - Push data storage back to the user/organisation?
  - Scope any central service
- Need to think who you collect this from...
  - End-users @ point of use: Hard to define parent organisation to collect money from
  - Projects: Hard to link user activity to a project. And to someone who can pay!
  - Organisations: Easiest people to invoice! But, like to know who they are supporting

# Data Models

- Evaluate Upfront Costs:
  - Given increasing performance in storage (PB/\$) most costs in first few years
  - However, migration costs are non-trivial (25% CERN tape capacity migration)
- Distributed Storage:
  - Can local storage be relied on for sustainability? Replicate for resilience?
  - Networking has improved (Gb/\$) and is 'free' at a European level (but not locally!)

# Conclusions

- Need to have an optimisation goal in mind
  - Define a budget, researchers always want something for nothing
  - Avoid the tragedy of the commons, define a cost for all services
- Sustainability is not another EC project
  - EC projects are fine for development, support, expansion, ...
  - Look to the community to provide what is essential to the community