What Can We Learn from15 Years of Sequencing Data Management?

A Reflective Case Study from a Bioinformatics Core Facility

Peter Briggs ***Genomic Bioinformatics Unit, Faculty of Biology Medicine & Health, University of Manchester UK***
peter.briggs@manchester.ac.uk

The University of Manchester

Introduction

The Genomic Bioinformatics Unit (GBU) is part of a University of Manchester core facility, providing processing and analysis of biological sequencing data and storing these data on behalf of researchers.

This poster is based around a personal record of data management activities within the facility from 2011 - when the data store consisted of 14 sequencing runs (~5 TB) - to the present day, where it now contains data from over 1,200 runs (~280 TB).

Analysing the historical record

A historical record of operational data management activities from 2011 to the present was reconstructed from operational notebooks and other facility records, and a qualitative analysis was performed with help from Microsoft Copilot.

This record is necessarily incomplete and reflects a personal operational perspective. Nevertheless, we believe that the recurring patterns and themes identified here still have value, and that taking a data stewardship perspective can yield further useful insights.

A simplified timeline: key events & turning points

2011
Starting point: automation & organisation begin
Scripts standardise operations previously performed manually

2012
Storage capacity first becomes a problem as data grow
Relocation and consolidation require validation & verification

2012
New sequencing technology changes our workflow
Processing data from Illumina sequencers drives automation

2013–14
Scale creates capacity issues & prompts mitigations
More runs = more data & requires slimming and relocation

2014
Operational metadata become more explicit
Unique IDs & structured metadata implemented to support operations

2014
Storage capacity prompts first archiving attempt
Exposes incomplete metadata and lack of suitable archiving tools

2015
Metadata moves alongside the data
Metadata stored next to datasets supports reporting and processing

2016
Capacity issues force further remediation of data storage
Relocating data store increasingly demanding & time consuming

2017
New experiment types stretch existing models
Single-cell data requires new processing and metadata

2017
Demonstration of value of long-term data retention
Researcher requests several datasets 5 years after generation

2021–23
Archival tooling emerges from capacity management
New tools implemented as archiving adopted to reduce storage footprint

2024
Another new sequencing technology disrupts our model
Handling ONT PromethION data requires revised code and practices

2024–25
Long-term archiving becomes operational
Older data systematically archived into cold-storage staging

2025
Archiving effort exposes fragmented metadata storage
Complete metadata for archive is spread over multiple locations

2025–26
Recovery and reuse expose wider questions
Recovery of data raises questions of access, ownership & governance

What can we see in a high-level view?

Recurring patterns

- Storage pressure
- Changing technologies
- Growing metadata requirements
- Continuing value of old datasets

Reactive adaptations

- Automation efforts
- Expanding metadata capture
- Remediation and archiving

Themes begin to emerge

- Practical adaptation worked!
- Automation also enabled standardisation
- Quick solutions contributed to future complexity
- File management broadened into data management

Connections are revealed

- Storage decisions affected archiving
- Archiving exposed metadata requirements
- Dataset recovery exposed ownership and access questions
- Retention incurred costs but also enabled later reuse

Reflections: from operations to stewardship

Could explicitly approaching these connected concerns as data stewardship change how we respond to them in future?

Define the data lifecycle

Making the implicit data lifecycle explicit could inform future decisions about processes and infrastructure

Metadata as a first-class asset

Treat metadata as an asset in its own right, rather than just as something only needed for specific facility operations

Formally recognise custodianship role

Recognition of the facility's custodial role could clarify decisions around access, reuse and retention of data with researchers and the broader institution

Use automation to embed sustainable practice

Automation could be used more deliberately as a tool to purposefully strengthen agreed standards and practices

Final takeaways

- **Reactive adaptation worked:** it enabled the facility to accommodate substantial technological, operational and organisational change.
- **Looking back changed our perspective:** issues encountered separately over 15 years appear increasingly connected, while some questions extend beyond technical solutions.
- **Stewardship offers a possible next perspective, not a proven solution:** other perspectives might offer different insights.

An open question: could making stewardship explicit change the way we respond to future challenges?