RDMF18
Transforming Data Storage for Life Sciences Research

“Managing research data through its lifecycle, and across multiple services”

Paul W Jeffreys
21 November 2017
Our mission is to make the discoveries that defeat cancer.
Presentation Outline

**ICR’s Lifecycle Management Requirement for Research Data**

1. Institute of Cancer Research / Royal Marsden
2. ICR Research Infrastructure
3. RDS Service: solution, rationale
4. ICR Research Data Storage Programme

**Which Metadata Catalogue Service?**
1. Institute of Cancer Research and The Royal Marsden Hospital
Institute of Cancer Research - at a glance

- Top 4 global cancer research organisation
- Top-ranked UK academic institution (REF)
- 20 drug candidates discovered since 2005
- More than 1,000 staff
- £161.9m income
- £110.0m expenditure

- Awarded Athena SWAN Silver
- More than 900 scientific papers
- Partnerships with 163 different companies
- Top UK university for invention income
- 141 research students
- 143 MSc students
ICR Academic Successes

The ICR is ranked as the top academic research centre in the UK; first in the *Times Higher Education* league table of university research quality compiled from the Research Excellence Framework (REF 2014)

Joint top of the *Times Higher Education* table for Innovation – based on worldwide citation of research in patents

Joint top of the *Times Higher Education* table for Innovation – based on worldwide citation of research in patents
Substantial and diverse funding

2016

- Total income £162m
- HEFCE 12% based on research excellence
- Grant income 38%
- Legacies and donations 7%
- Invention income from our discoveries 16%

-> Services not centrally funded

Total incoming resources 2016

- 12% Higher Education Funding Council for England
- 38% Research grants
- 16% Royalty income
- 7% Legacies and donations
- 3% Investment and tuition fees
- 24% Sale of part of our future royalty stream
ICR Research Divisions

- Breast Cancer Research
- Cancer Therapeutics
- Genetics and Epidemiology
- Radiotherapy and Imaging
- Cancer Biology
- Clinical Studies
- Molecular Pathology
- Structural Biology
Partnership with The Royal Marsden, and “bench-to-bedside and back” approach:
-> discoveries made, and clinical impact delivered, uniquely

Outstanding record of research achievement dating back more than 100 years
ICR Knowledge Hub

- Big data
- Artificial intelligence
- Machine learning
- Identifies treatment algorithms

Collect all data
- Genomic, proteomic & other molecular profiling
- Histopathology
- Clinical imaging
- Clinician- and patient-reported data
- Therapy information
- Primary and secondary care data

Recommend drug-drug or drug-radiationtherapy combo
- Individual patient
- Specific time

Treat

Future – dynamic adaptive therapy

Plan to start novel dynamic adaptive individualized trials from 2022
- Drug-drug in advanced prostate (de Bono); lung (Swanton)
- Drug-radiotherapy in lung, prostate, breast (Harrington et al)

Improved Outcomes
Making the discoveries
Our strategy to defeat cancer

The ICR and The Royal Marsden delivered a joint strategy covering the next five years

Our vision
We will overcome the challenges posed by cancer’s complexity, adaptability and evolution through scientific and clinical excellence, innovation and partnership

First pillar: Unravelling cancer’s complexity
Comprehend the full complexity of cancer by harnessing the power of new technologies and Big data
The London Cancer Hub
A global centre for cancer innovation
http://www.icr.ac.uk/our-research/our-research-centres/london-cancer-hub

To deliver: an exceptional environment for cancer research that enhances the discovery of new treatments and their development for patients

To provide: state-of-the-art facilities, and be joined by a multitude of high-tech enterprises in a network of 10,000 researchers, clinical staff and support staff all operating from one site
2. ICR Research Infrastructure
HPC Infrastructure:

HPC Clusters

- c 1,800 cores,
- 12-16GB per core,
- 2PB scratch storage
- Designed for a parallel workload

- ~70% average use (enough to keep ahead of the instruments)

Both HPC and storage have been dominated by NGS; new imaging systems are changing the balance (4TB/hour)

We know nothing about the incoming data other than username and some educated guesswork; no direct information about what happens to the data going through HPC either
Existing Research Infrastructure:

Infrastructure mainly at offsite datacentre – highly secure and resilient

Storage system geographically distributed between London and Slough

All three sites are connected by a private 10Gb network (with an option to increase b/w)

Traffic will reroute in the event of failure

Each site has a 10Gb internet (JANET) connection
RDS Service Network
3. RDS Service: solution, rationale
RDS Service - User Requirements

Collected from researcher workshops:

- 6PiB* storage, expandable to growing research need; minimum 20PiB capacity
- High level redundancy ensuring robust solution
- Cost effective and competitively priced solution (-> vendors offered two Tiers)
  - Rapid access to data held in Tiers
  - Ability for researchers to manage data transfers between Tiers
  - Single namespace
  - Direct access to data in Tier 2
- Ability to protect against accidental loss of datasets
- Staging Areas to guarantee access to storage from instruments

* PiB is a binary unit of storage equivalent to 1.13PB (decimal version)
Architecture

Data source

Transient Storage

Main RDS Storage

Research Data Storage (RDS)
Research Data Storage Project

Project Overview

- Delivered:
  - Tier 1: December 2016
  - Tier 2: May 2017

- Total user capacity > 6.5 Petabyte
- Current use: c. 1 Petabyte stored

- Some data still to be transferred from previous storage service

- Completed in August 2017

RDS: two Tiers connected by a WOS Bridge
Architecture

Tier 1 is a stretched cluster with synchronous mirroring between Sutton and Slough; total usable capacity of 2PiB

WOS Bridge Servers at both sites for migration and retrieval of data at both sites eg. an HSM implementation

Tier 2 is a WOS object store (three zone GOA configuration); space and power means that Chelsea could not be used; 4PiB usable capacity
Management of Data

ICR has ~140 research teams; each allocated a Tier 1 fileset

The directory tree is laid out exactly as in Active Directory structure is (same applies to scratch disks on HPC):

  Division > Unit > Team > Person > …

- Each fileset is has a snapshot taken every day
- Snapshots are retained for 90 days by default
  - The snapshot also includes the OID for any files which have been stubbed and the data migrated to Tier2

- Each fileset has a quota on it; individuals may also have a quota
Management of Data

• Data can be migrated from Tier 1 to Tier 2 based on Bridge Rules
• Default rule: files which have not been accessed for 90 days to be migrated
• Most files migrated to WOS are managed by the Global Object Assure policy
• Smaller files (<10MB) on WOS are managed by simple replication between Sutton and Slough – more efficient
• Plan to implement a user-controlled “no-migrate” option for their data using extended attributes shortly
• Usage of Tier2 cannot be quota’d but can be measured
4. ICR Research Data Storage Programme
Research Data Storage Programme:

• RDS Service at core of RDS Programme
• Projects to develop two new components; also Dropbox Business service
RDS Programme Vision

RDS Programme

- Support world-leading research for ICR and The RM
- Life cycle management for ICR research data (*nb ICR culture*)
  - Capture information at creation - Long term storage
- Simple and efficient to use; no appreciable overhead to researchers
- Cost-effective
- Scalable
- Highly resilient
- In future: underpin ‘big data analytics’ capability
  - ICR Data Science department established
Questions for RDMF18 Attendees:

Does your institute/university:

1. Have a single central live data storage service (used by majority of researchers)?
2. Plan to have a single central live data storage service?
3. Currently have a central storage service > 20Petabyte?
4. Plan to have a central storage service > 20Petabyte within 2 years?
5. Have a single name space across central storage?
6. Plan to have a single name space across central storage?
7. Have a repository for long term storage/curation of research data?
8. Plan to have a repository for long term storage/curation of research data?
9. Do you deliver Research Data Management (RDM), making data held searchable?
10. If you have deployed Metadata Catalogue: which, how much resource required?
Scientific Computing Research Data Strategy

Research Data Strategy Outline:

• Address managing lifecycle of research data:
  • From: STiMS, Staging Areas, RDS Service, External Data Sharing, ICR Business Dropbox Service, Repository

• Lifecycle coordination:
  • Cover data movement as it passes across and between services provided by SC
  • Address Information Governance across this data lifecycle
  • Consider how plan might serve to underpin a broader set of activities with research data
    (machine learning, big data analytics…)
Example workflow:

- ICR Instruments
- Lustre (1.5)
- External Data Sharing
- Sync & Share

Staging Areas 8x30TB

NAS Layer (NFS & SMB)

Tier 1 (2 PiB)

Tier 2 (4 PiB)

Repository / Archive
Likely minimum Meta Data required for each dataset:

- Individual owner, team owner, when created (and by whom), instrumentation setup, how long needed
- Description (textual description)
- Data classification (public, local, restricted)
  - Patient data; anonymised, consented, ....
- State of data (live, no longer active, publishable, embargoed...)
- Provenance
  - Which services where held and when
  - Which applications used
- Hierarchy (parent, child)
- Workflow details eg:
  - When data arrives on staging area - do this
  - Processing pipeline information, applications version

*What have we missed; given that very difficult to collect data?*
Metadata Catalogue solution?

We think we have relatively simple requirements – to add metadata and generate workflows: *is this naïve?*

*Many potential solutions: ICAT, iRods, Dspace & other library-based services, SRB, Twist, …*

• Seem to be either designed for facilities (beam lines..) or library services
• Most seem to be heavily tailored for local use

*Looking for solution that is:*
• easy for scientists to use
• largely off-the-shelf, and able to deliver PoC in one month of one person

*What would you recommend please?*
Questions

paul.jeffreys@icr.ac.uk