UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK

We will keep it forever

(from a tin perspective)

Marc O'Brien & Peter Maccallum

IT & SC Department

CRUK Cambridge Institute

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK

CRUK Cambridge Institute

The CRUK Cambridge Institute generates genome, multi-modality imaging and histopathology data, and is rapidly approaching 2PB of stored data. Here we discuss a scientist driven, ‘keep forever’ approach to data that supports publications.

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK

Data Lifecycle

- Because data are valued assets, we need to manage data over their entire lifecycle beyond the immediate need.

- The goal of managing over the data lifecycle is to eliminate waste, operate efficiently, and practice good data stewardship.

Courtesy of United States Geological Survey
http://www.usgs.gov/datamanagement/why-dm/lifecycleoverview.php

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK

A Loose Definition of Scientific Data

Scientific Data

- Raw data from acquisition machines.
- Papers you are working on.
- Reference data.
- Processed data.
- Intermediate analysis data.
- Experimental results.

Not Scientific Data

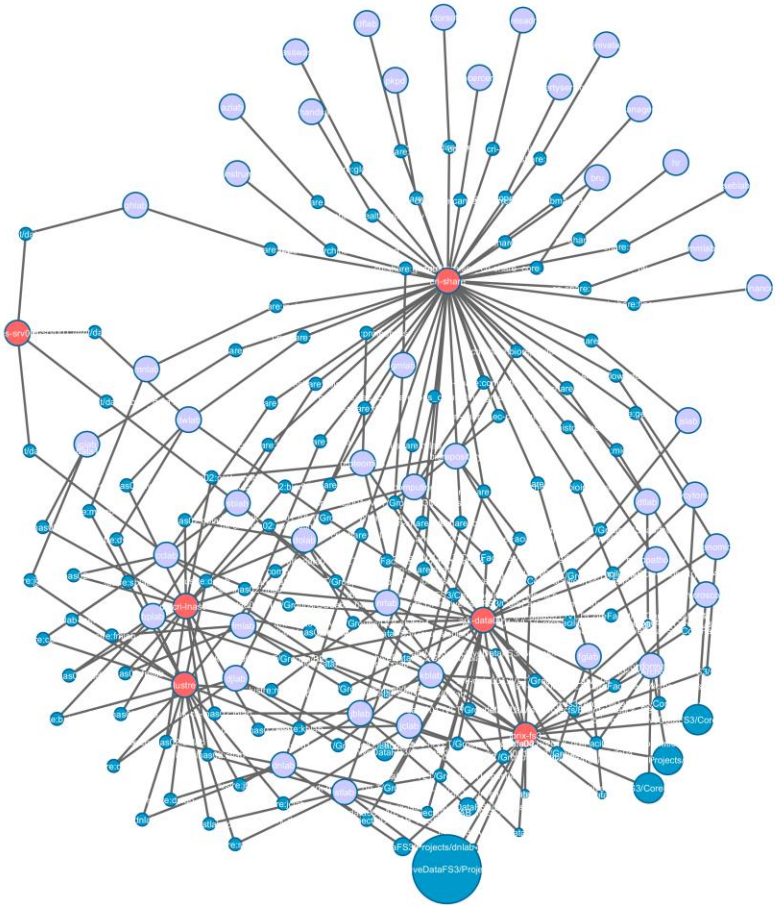
- Holiday snaps.
- Music collection.
- Computer desktop.

…whatever is produced in research or evidences its outputs (Marieke Guy, UKOLN University of Bath)

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK

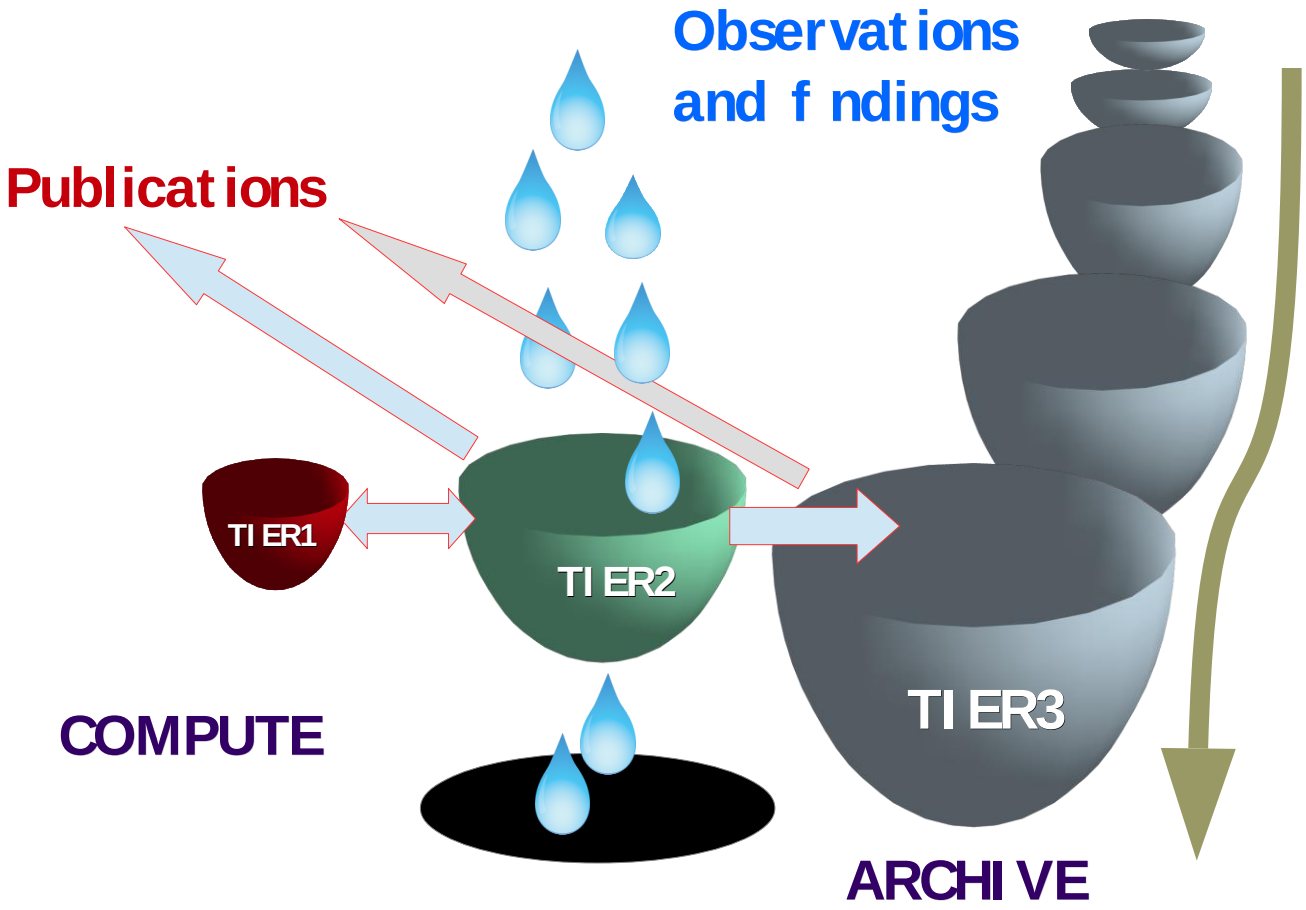
Institute File Systems and Folders



UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK

The Institutes Tiered Storage



UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK

Data currently stored at the Institute

- Tier1 (Scratch)

  ~70TB
- Tier2 (Working Data, Replicated)

  ~246TB
- Tier3 (Archive)

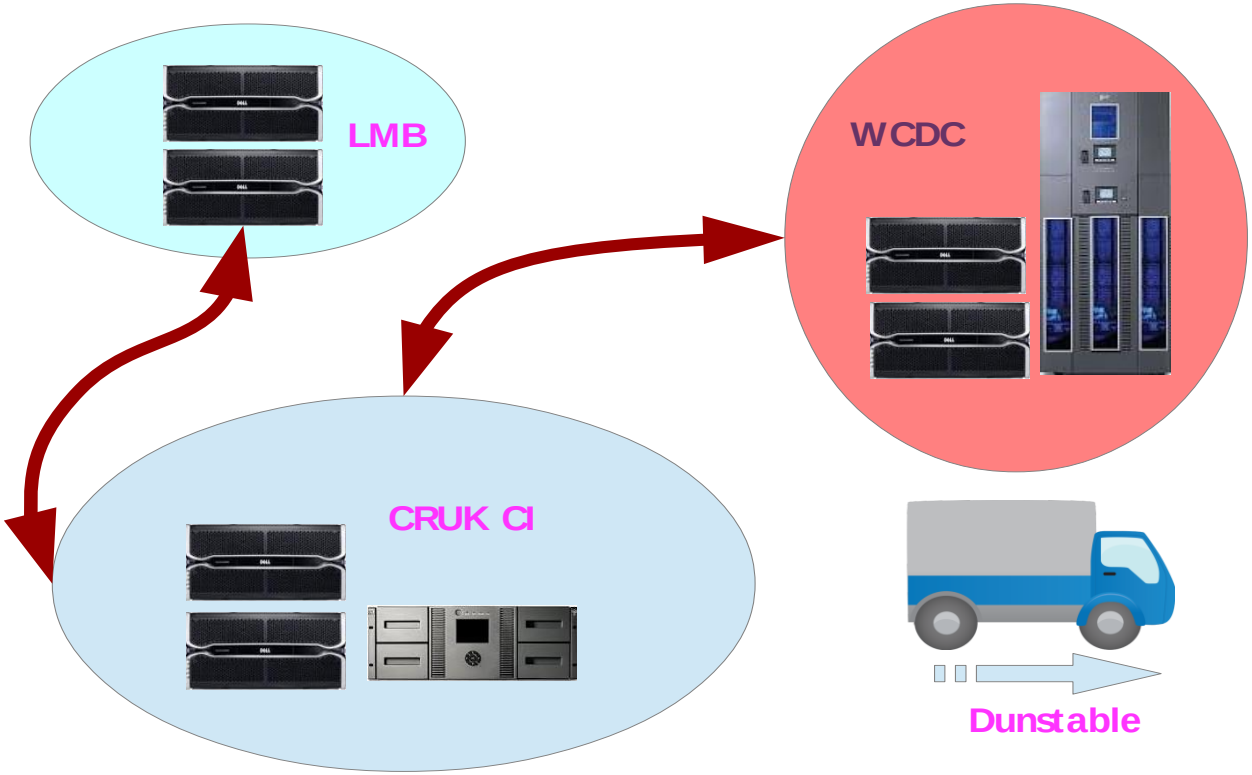
  ~1.2PB

Note: This is actual data files, file systems total > 2PB

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK

Data Replication for Tier 2 and Tier 3 Archive tape backups



UNIVERSITY OF CAMBRIDGE

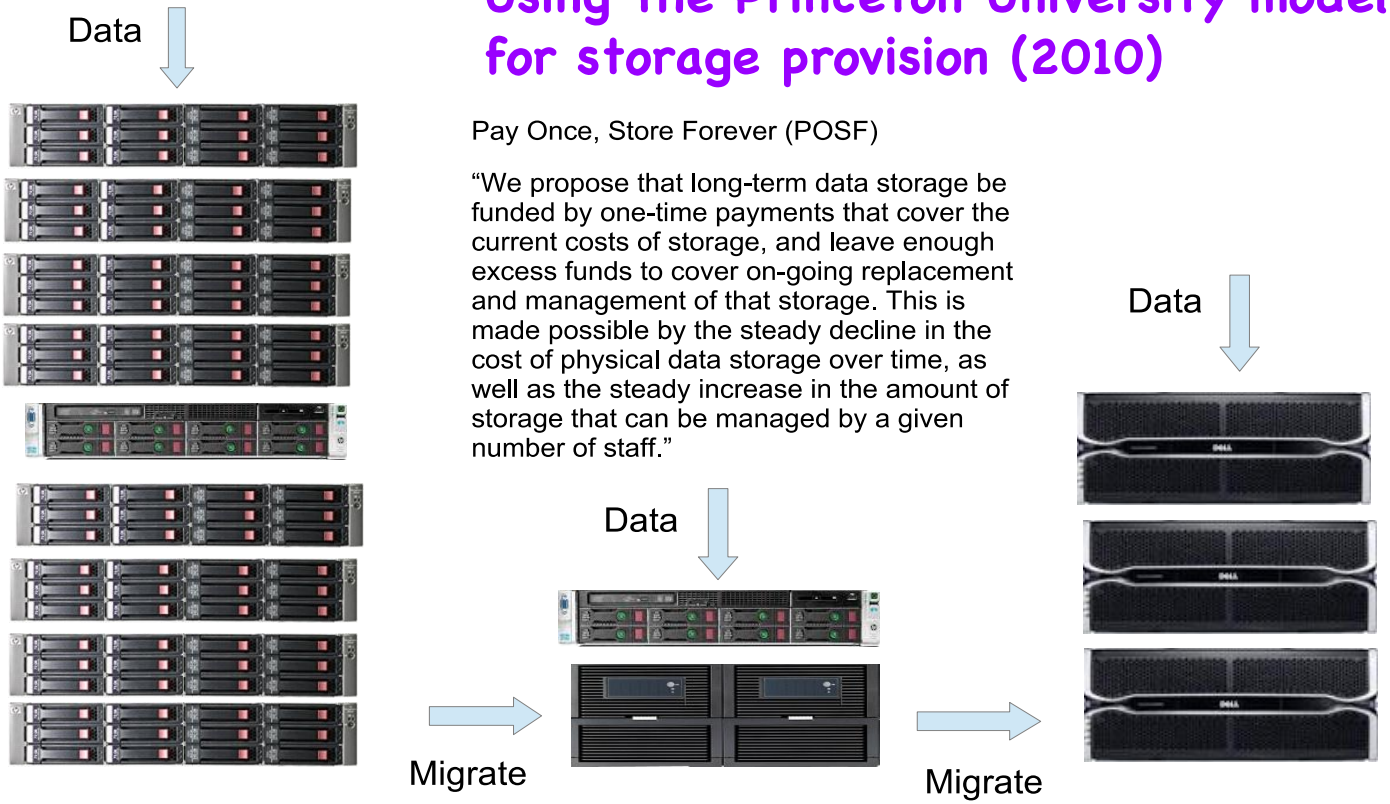
CANCER RESEARCH UK

Princeton

Using the Princeton University model for storage provision (2010)

Pay Once, Store Forever (POSF)

"We propose that long-term data storage be funded by one-time payments that cover the current costs of storage, and leave enough excess funds to cover on-going replacement and management of that storage. This is made possible by the steady decline in the cost of physical data storage over time, as well as the steady increase in the amount of storage that can be managed by a given number of staff."



~ **2007** 540GB drives
42TB in 18U

~ **2011** 2TB drives
104TB in 7U

2016 8TB drives
1080TB in 12U

http://arks.princeton.edu/ark:/88435/dsp01w6634361k

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK

Tips for implementing Princeton

- Buy storage hardware with ‘x’ years warranty.
- Set an ‘x’ year storage hardware refresh cycle.
- Include data migration as part of any storage procurement project.
- Be strict with quotas and data ownership.
- Advise on ad-hoc storage purchases.
- Monitor acquisition technology changes.

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK

The Quota Collie and the Data Sheep

https://upload.wikimedia.org/wikipedia/commons/thumb/c/ca/BC_eye.jpg/300px-BC_eye.jpg
C. MacMillan - Original Work
BC_eye.jpg, Border Collie exhibiting "Collie Eye" to stare down sheep
Permission details
CC-BY-2.5

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK

Managing the Data Flow

- Tier 1 is transitory, results are copied off and files deleted when processing is finish.
- The Tier 2 file systems are of semi fixed size, once a project/experiment is finished, a subset of the files is archived and the rest deleted. Tier 2 size increases with technology step changes or when new research groups join the Institute.
- Only scientific data should be archived, archived data will be kept forever….
- **All of this is driven by the quota sheepdogs ☺**

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK

Tier 2 Working Data Backup Retention Policy

- Tier 2 is replicated to offsite storage.
- Tier 2 storage is RAID 6 protected.
- Rolling Daily, Weekly and 3 x Monthly snapshots to the secondary storage.
- Monthly tape backup from oldest monthly snapshot.
- 3 month tape retention (within library).

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK

Tier 3 Archive

- Modular Write Once Read Many (WORM) archive system.
- Archive storage is Dynamic Data Pool equivalent of RAID 6.
- Per group/core facility file systems.
- Each file system can grow to 64TB.
- Files written have checksums stored as separate metadata.
- Files written are copied to tape once.
- Tape backups are stored offsite at Dunstable.
- Tapes are migrated as new densities become available.
- Data stored is kept forever and migrated to new hardware when appropriate .
- Single namespace presentation.

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK

The Move Towards Open Data

- Royal Society report, June 2012

Universities have a major role to play in supporting open data.

- Finch Report, June 2012

Measures in the UK to encourage the further development and use of repositories could lead to significant improvements in access to publications and reports arising from UK research. The benefits would be perceived within universities in facilitating research management, in providing a showcase for research outputs and expertise, and in providing a mechanism for the management of research data.

- RCUK, July 2012

Open Access Policy on peer reviewed publications adopted by EPSRC and AHRC (Primarily relates to journal choices).

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK

Cambridge University Supports Open Data

- Data Management Guide

  http://www.data.cam.ac.uk/data-management-guide

- Research Data Management Policy Framework

  http://www.data.cam.ac.uk/university-policy

- Repository

  https://www.repository.cam.ac.uk

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK

Work in Progress

- Investigation of metadata layers and object stores for the purpose of routinely exposing the published and supporting data that we already store.
- Implementation of an Institute data preservation and sharing policy with guidelines for project folder naming conventions etc.
- Tectonic plate replica separation….

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK

THANK YOU ALL