Data Management Workshop
Cape Town 3-6 June 2014

H3ABioNet

Pan African Bioinformatics Network for H3Africa

Data Archive Solution

Outline

- Archive project background
- H3Africa data archive requirements
- Data archive needs analysis
- Data archive solutions investigated
- Current project status and next step
- Data submission process
- Data access process
- Questions

Archive Project Background

- H3Africa directive to develop a data archive
- The Infrastructure Working Group (ISWG) took over all responsibility for this project
- Data Management Taskforce (DMTF) created to investigate and develop a data archive solution

H3Africa Data Archive Requirements

- Design and implement a data archive solution to house a copy of the H3Africa research data in Africa
- The archive needs to be secure and reliable
- Research data should be held for a maximum of 9 months before being submitted to EGA
- Assist nodes with EGA compliance
- Assist nodes with EGA submission
- Retain data until the project ends in 2017

Data Archive Needs Analysis

- How much disk space is required?
- Disaster Recovery?
- How would we transfer data?
- How would we process data?
- What data is stored on the archive?
- How would we secure the data?

Needs Analysis Summary

- In summary we needed a solution with:
  - A minimum of 250TB of disk space
  - Implement disaster recovery measures geared towards a data archive
  - Include a data staging area
  - Be flexible enough to accommodate various data transfer mechanisms
  - Create a searchable metadata database
  - Solution needs to be secure and reliable

H3ABioNet Option

- Pros
  - H3ABioNet has full control of the data archive solution and the network infrastructure
- Cons
  - Purchase all hardware reduces overall buying capacity
  - Responsible for power, cooling and hardware refresh costs
  - Responsible for any and all DR measures

SAGrid Option

- Pros
  - Access to data centres and network infrastructures
  - H3ABioNet retains ownership of the data archive hardware
  - Access to larger support base
  - Gain access to additional value added resources in Africa
    - Alternative transfer mechanisms
    - Federated authentication
- Cons
  - Reduced buying capacity
  - Value added resources still in development

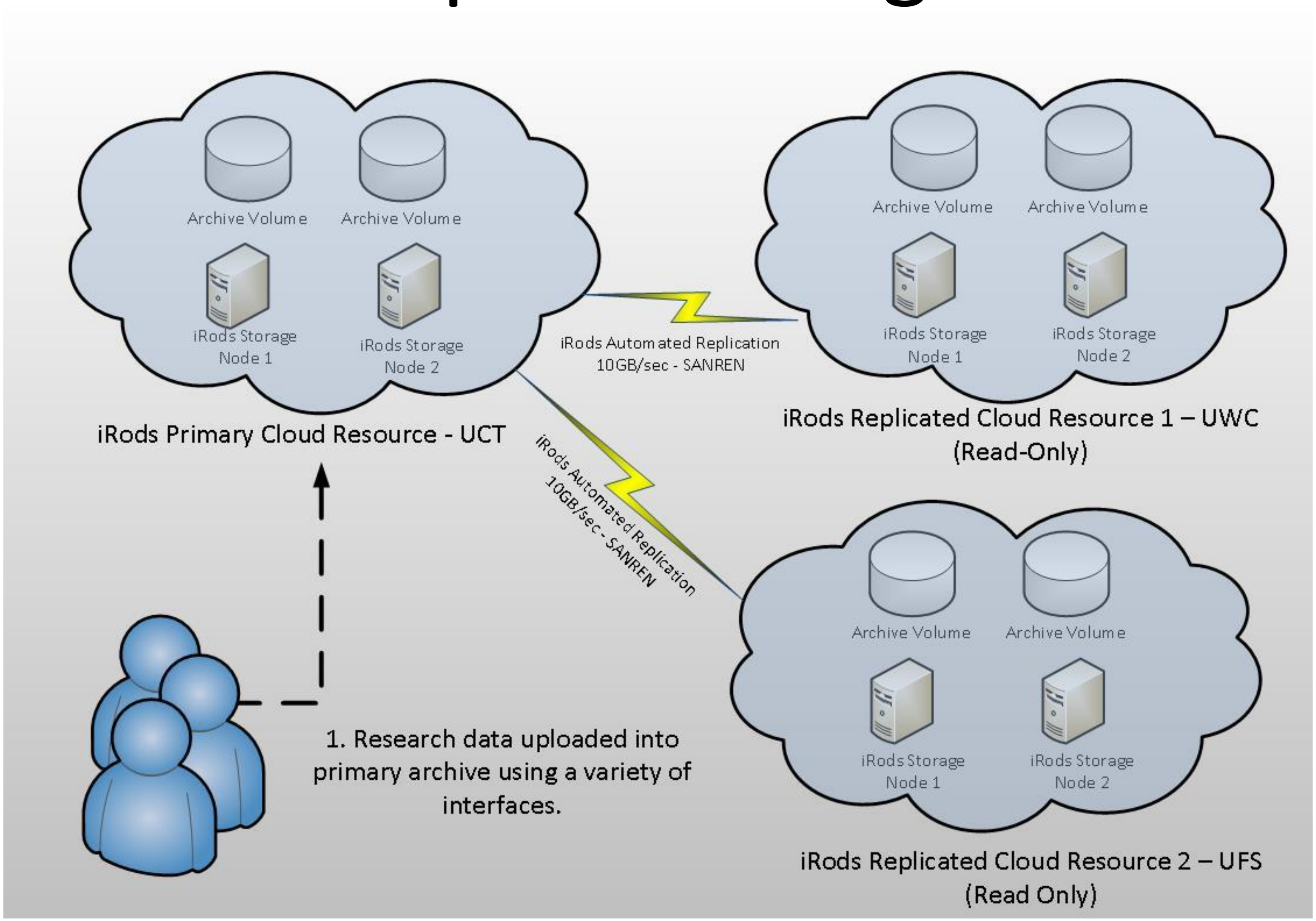
UCT Option

- Pros
  - UCT will manage and maintain the backend data archive infrastructure (bandwidth, physical disks, virtual servers, etc)
  - Replicate data across two institutions to increase data recoverability
  - Provide 500TB of archive disk space
  - Absorbs the secondary cost of electricity, cooling, etc
  - Willing to retain the data archived to disk for 10 years post 2017
  - Incur hardware refresh cost (3 year cycle)
- Cons
  - H3ABioNet does not own the data archive hardware –only the content
  - Will have to log a support ticket to resolve faults which will impact on turnaround times

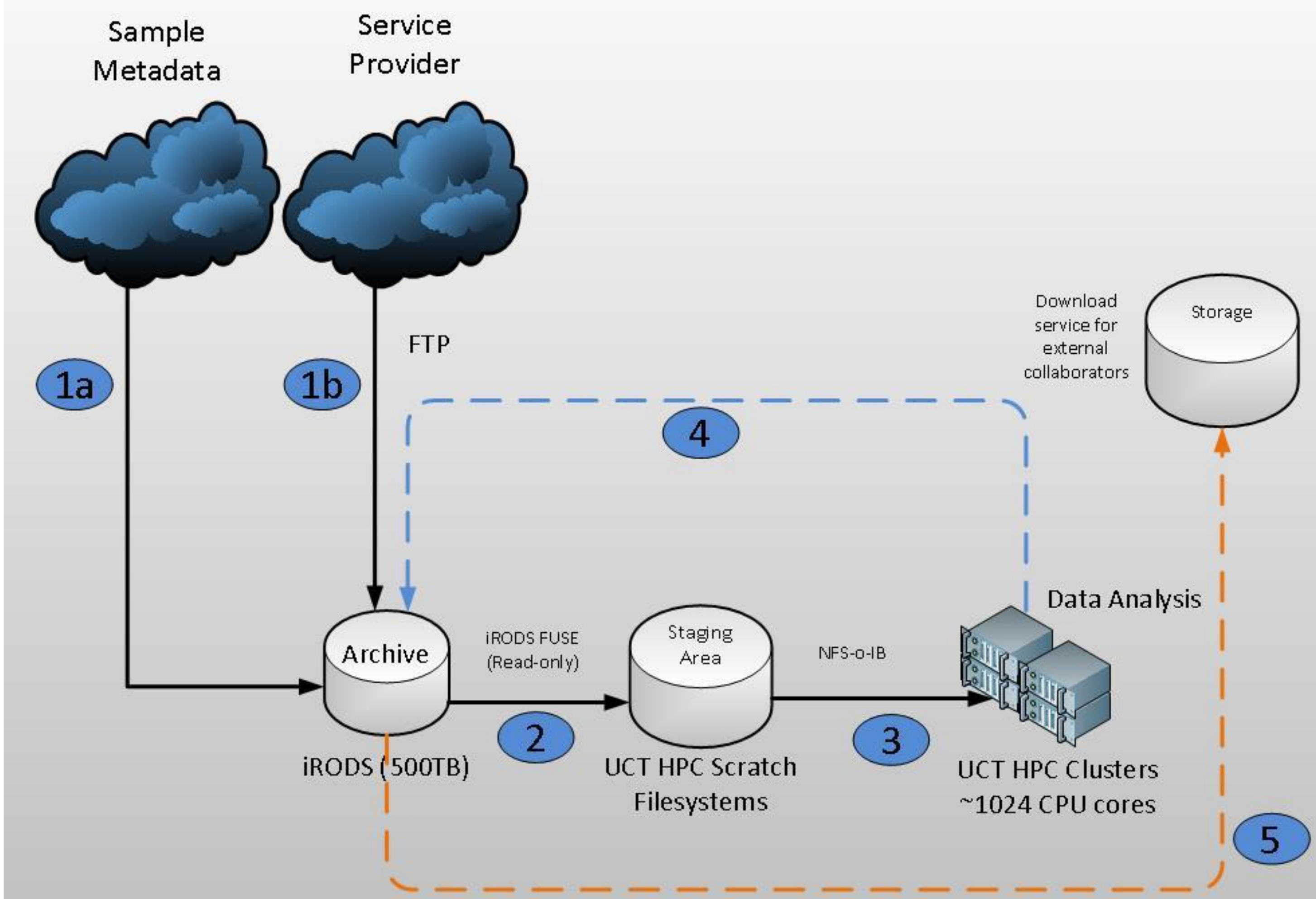
Current Project Status and Next Step

- Based on the options investigated, UCT made the best sense as the hardware solution provider
- SAGrid collaboration
- Project split into three phases:
  - Phase 1 – provision servers, disk space and initial transfer mechanism
    - Currently in the proof of concept stage
    - Transferring data via Globus Online
    - Procuring physical storage for the data staging area
  - Phase 2 – develop metadata database
  - Phase 3 – integration with SAGrid

Proposed Design



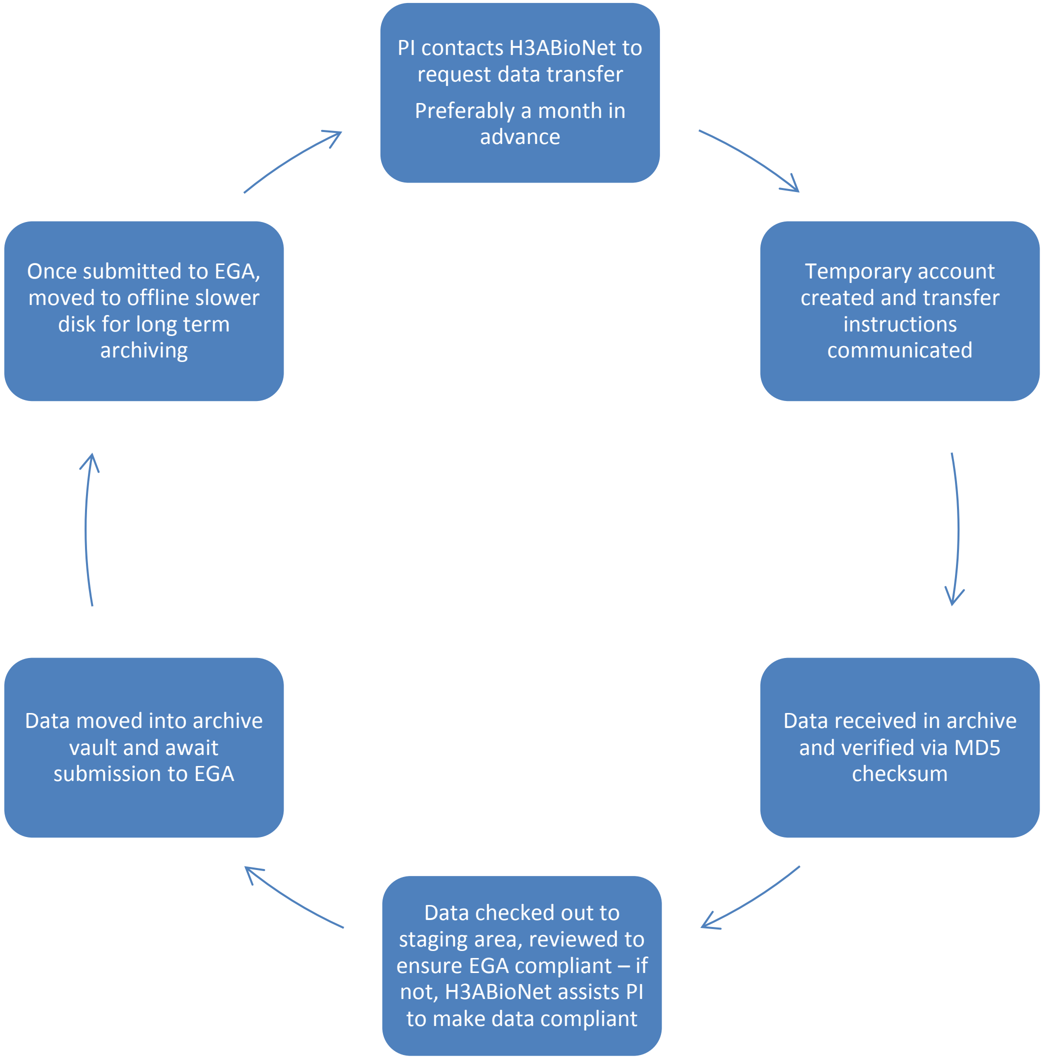
Data Flow


Sample Metadata
Service Provider
FTP
1a
1b
4
Download service for external collaborators
Storage
Data Analysis
Archive
iRODS FUSE (Read-only)
Staging Area
NFS-o-IB
2
3
iRODS (500TB)
UCT HPC Scratch Filesystems
UCT HPC Clusters ~1024 CPU cores
5

Data Submission & Access Policy

- Disclaimer
  - H3ABioNet does not own the data
  - The archive will not house the sole copy of data – PI's to retain a copy
  - H3ABioNet will only make data available based on the written authorization from H3Africa
  - H3ABioNet will not be responsible for data once downloaded
- Data Submission
  - Notify HDT a month in advance
  - Create a high-level folder with a H3ABioNet defined naming conventions
    - PI_shortdate_submission# = NicolaMulder_04062014_00
  - Supply readme file describing the data
  - Supply Data Submission Request Form
- Data Access
  - Access only granted based on DAR sent from H3Africa to HDT
    - Identify who needs access and to which data
    - No or incomplete DAR = No access

Data Submission Process



PI contacts H3ABioNet to request data transfer

Preferably a month in advance

Temporary account created and transfer instructions communicated

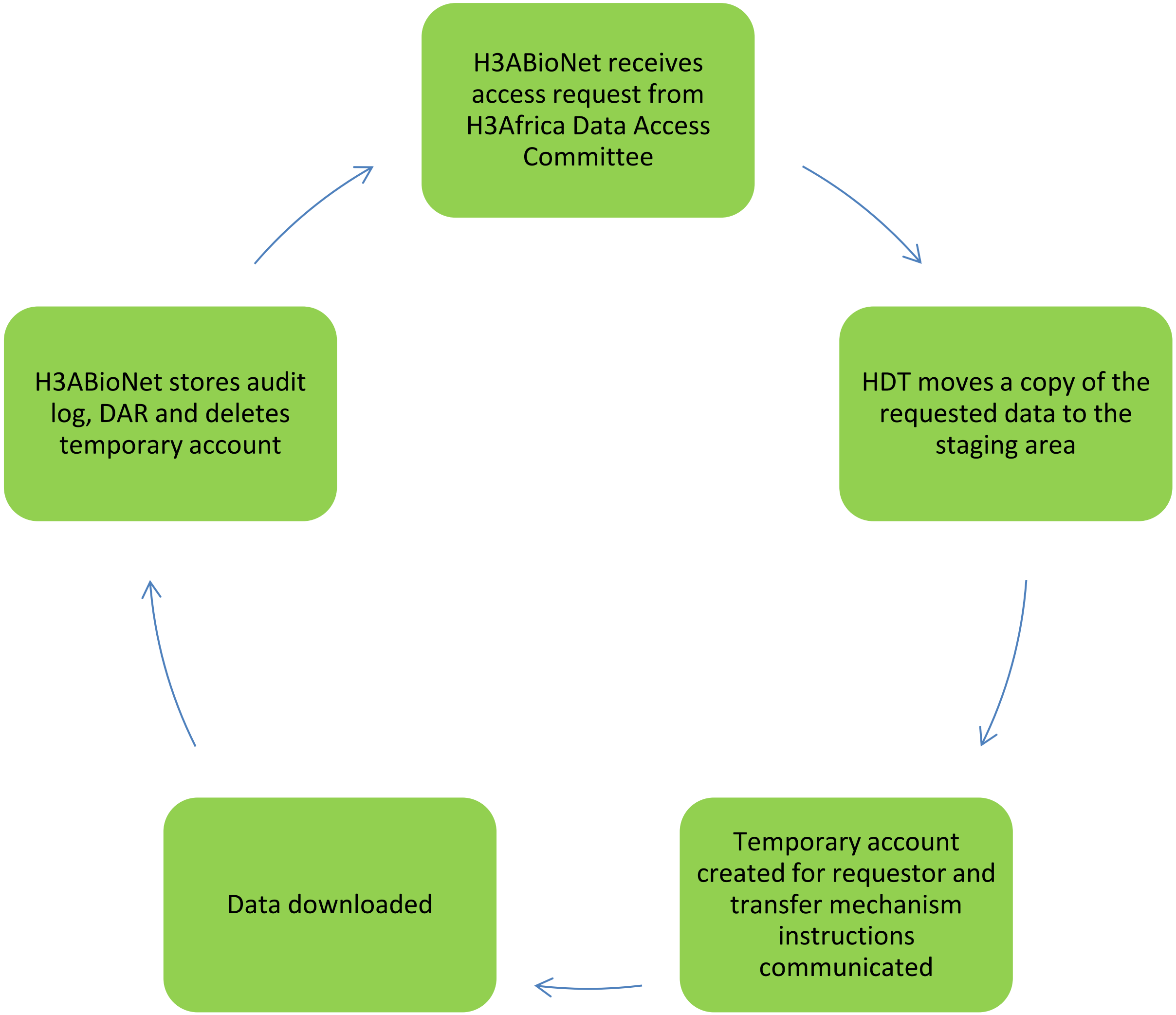
Data received in archive and verified via MD5 checksum

Data checked out to staging area, reviewed to ensure EGA compliant – if not, H3ABioNet assists PI to make data compliant

Data moved into archive vault and await submission to EGA

Once submitted to EGA, moved to offline slower disk for long term archiving

Data Access Process



H3ABioNet receives access request from H3Africa Data Access Committee

HDT moves a copy of the requested data to the staging area

Temporary account created for requestor and transfer mechanism instructions communicated

Data downloaded

H3ABioNet stores audit log, DAR and deletes temporary account

Acknowledgement
Data Management Taskforce members

Taskforce Member	Organization
Ayton Meintjies	University of Cape Town – CBIO
Gerrit Botha	University of Cape Town – CBIO
Liam Thompson	University of Witwatersrand
Luda Mainzer	University of illinois
Mohamed Alibi	Institute Pasteur of Tunis
Scott Hazelhurst	University of Witwatersrand
Sumir Panji	University of Cape Town – CBIO
Suresh Maslamoney	University of Cape Town – CBIO

Questions / Concerns

H3ABioNet
Pan African Bioinformatics Network for H3Africa