syngenta

Managing Next Generation Sequence Data at Syngenta with iRODS

Todd Moughamer
iRODS UGM 2018
Durham, NC
June 6 & 7, 2018

Classification: PUBLIC

Who we are

A leading agriculture company helping to improve global food security by enabling millions of farmers to make better use of available resources.

- World-class science and innovative crop solutions.
- 28,000 people in over 90 countries working to transform how crops are grown.
- Committed to rescuing land from degradation, enhancing biodiversity and revitalizing rural communities.

90
countries

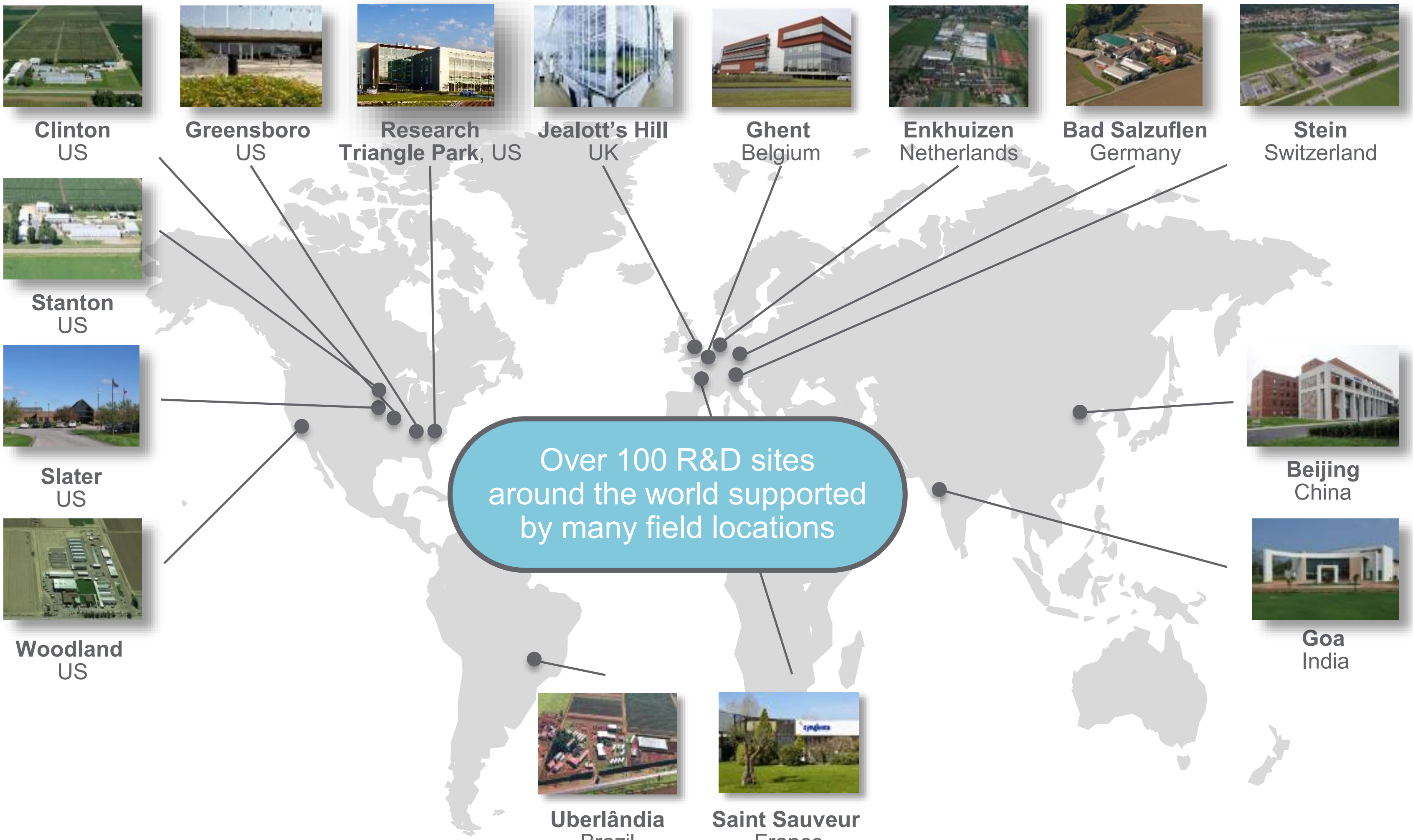
107
production and supply sites

119
research and development sites

27,810
employees

Key R&D centers across the world

Unrivalled global breadth



Clinton
US

Greensboro
US

Research
Triangle Park, US

Jealott's Hill
UK

Ghent
Belgium

Enkhuizen
Netherlands

Bad Salzuflen
Germany

Stein
Switzerland

Stanton
US

Slater
US

Woodland
US

Over 100 R&D sites around the world supported by many field locations

Beijing
China

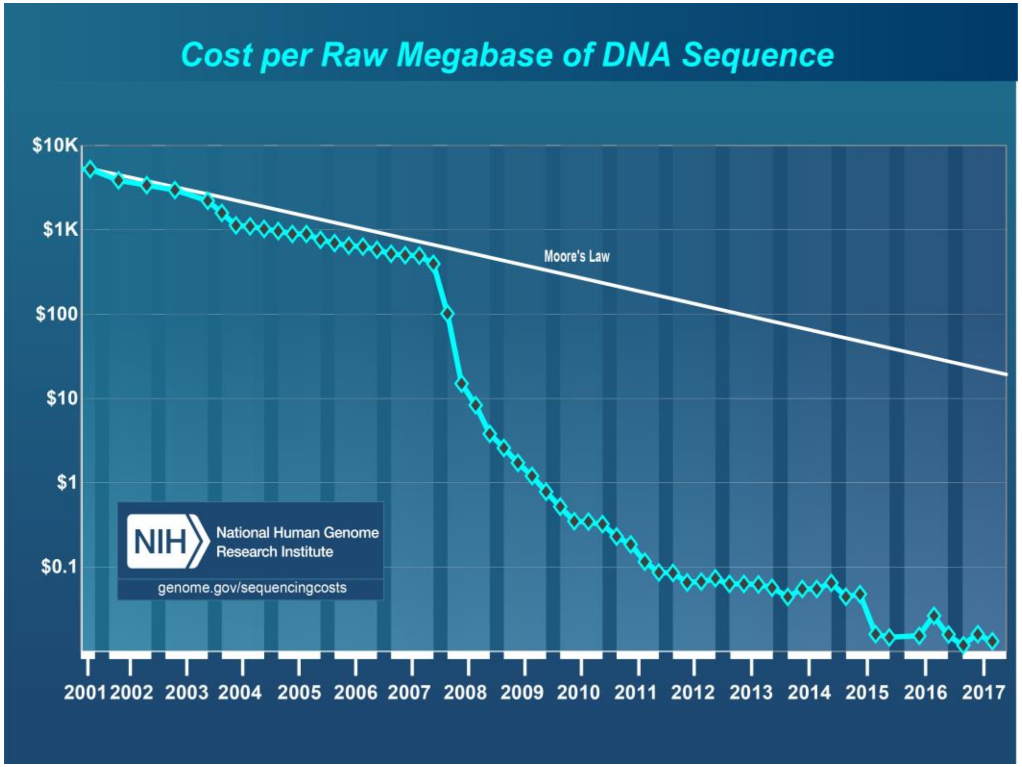
Goa
India

Uberlândia
Brazil

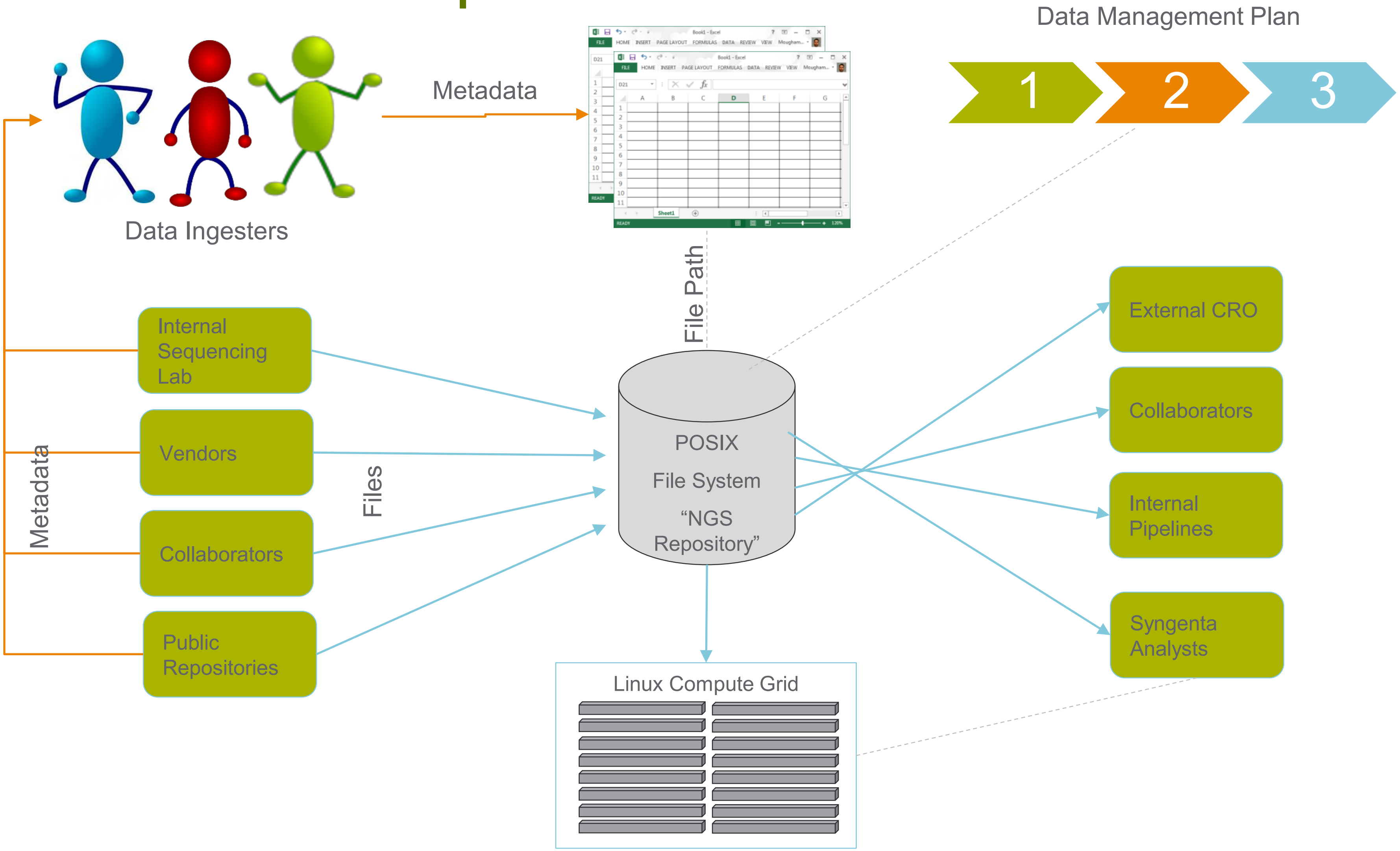
Saint Sauveur
France

Background

- Next Generation Sequence (NGS):
  - While the cost of DNA sequencing as been dropping for sometime, there has been a precipitous drop over the last ten years driven by the introduction of new technologies to the market



Our Situation: Simplified View



Data Management Plan

1

2

3

Data Ingesters

Metadata

File Path

Internal Sequencing Lab

Vendors

Collaborators

Public Repositories

Metadata

Files

POSIX

File System

“NGS Repository”

Linux Compute Grid

External CRO

Collaborators

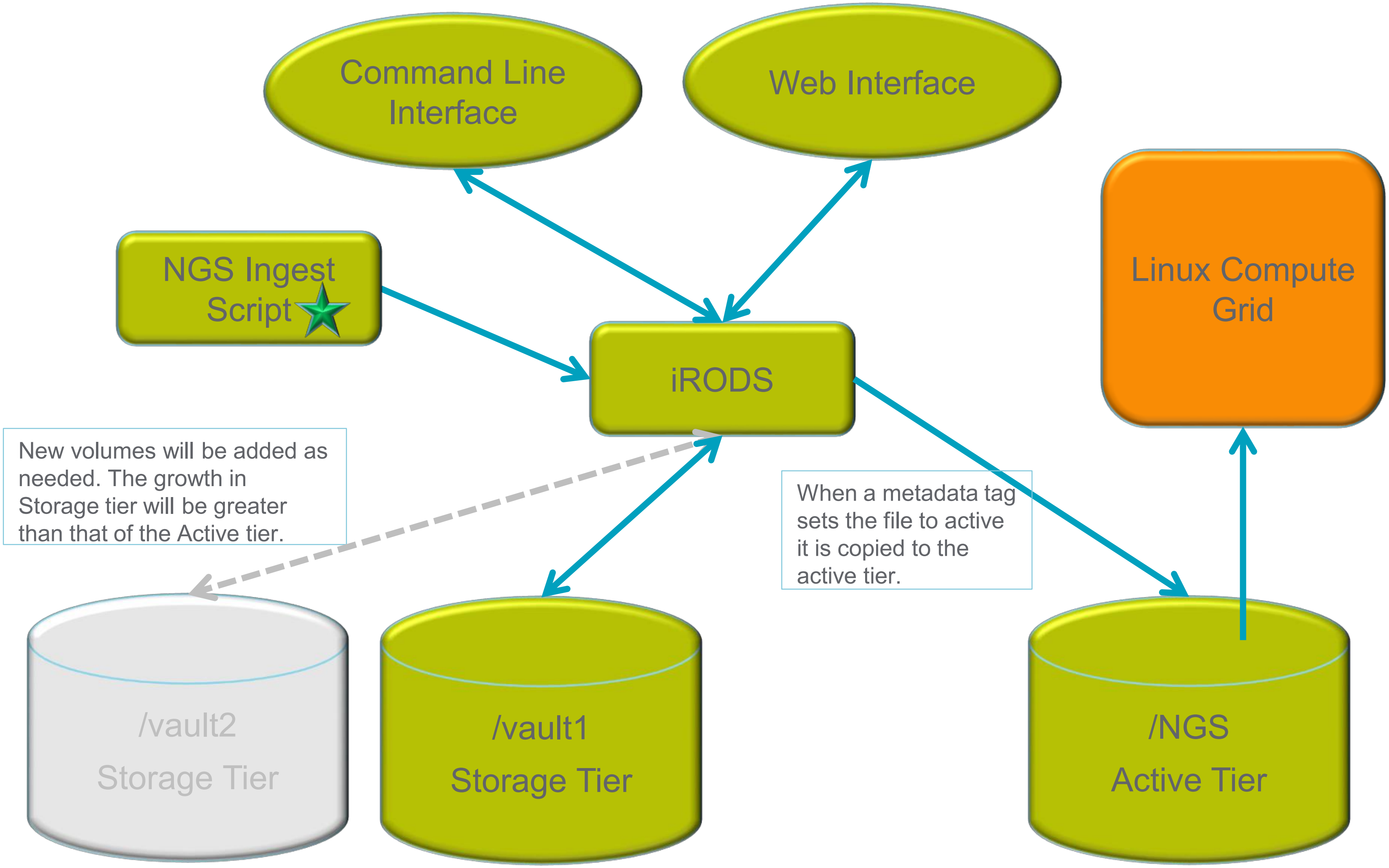
Internal Pipelines

Syngenta Analysts

Approach: iRODS Pilot

- Partner with external iRODS experts: iRODS Consortium at RENCI on set up and for consultation
- Focus on new data not legacy data
  - Our past attempts focused on legacy data but fell short when it came to adding new data
- Differentiate our active data (that which is to be analyzed) from data that simply being stored per our data management plan

iRODS NGS Pilot Implementation



New volumes will be added as needed. The growth in Storage tier will be greater than that of the Active tier.

When a metadata tag sets the file to active it is copied to the active tier.

★ Internally Developed

Implementation: Ingest Script

- Previous process: 1) Ingester copies data to repository filesystem, 2) records metadata into a spreadsheet, and 3) runs a script to verify & log the files added to the repository
- New process: Ingester executes one script to load data and metadata into iRODS and log the files that were added
- Input: Metadata spreadsheet with the file paths to the data
- Operation:
  - Validates metadata
  - Provides summary of what's to be loaded and prompts user to continue
  - Blocks object/file over-writes.
- Output:
  - Records files that are ingested in our existing central log file
  - Creates:
    - Log of its operation
    - Back-out scripts that can be used to remove the files or metadata from iRODS
- Details:
  - Python3 with Pandas via Anaconda
  - Calls iCommands

Implementation: Active Data Tier

- Previous process: Data that is being analyzed resides with data that is just being stored. No single source tracks what data is active or inactive. Moving data to a different volume is done by system administrators (rsync) and the metadata spreadsheets need to be manually updated with the new file locations.
- New process: Data ingesters can set a metadata AVU to cause iRODS to replicate data to our active volume. The AVU can be changed for iRODS to remove the data from the active volume. No need manually copy files or manually update spreadsheets.
- Input:
  - Added a new volume as an iRODS resource (ngs_active)
  - Configured rules to be triggered by attribute active_state
- Operation:
  - If active_state value set to 'true' for an object it is replicated to ngs_active resource
  - By default directories and files in ngs_active are set to be read by any Linux user
  - When active_state attribute is removed or set to value other than 'true' the file is removed from ngs_active
- Details:
  - iRODS Rule Language
  - msiExecCmd microservice to run chmod

Implementation: Restricted Access

- Previous process: When data is not to be accessible to just any user on the Linux system the system administrators limit access permissions to a specific user or group
- New process: Data ingester can provide a Linux group to iRODS through a metadata AVU. If iRODS copies that data to the active volume it will set the permission to allow read access to only that group.
- Input:
  - Operates with active tier
  - Configured rules to be triggered by attribute restricted_access
- Operation:
  - If restricted_access is set to ‘true’ with units set to a valid Linux group the permission of the files and directories on the active tier are changed from the default
  - Directories and files are set to only allow read access by members of the Linux group
  - When restricted_access attribute is removed or set to other than ‘true’ the permissions are changed to allow general read access
- Details:
  - iRODS Rule Language
  - ‘irods’ user must be a member of the Linux group…fails otherwise and defaults to group ‘irods’
  - msiExecCmd microservice to run chmod and check Linux group membership

iRODS Status

- Opened for new data ingest in November 2017 and has been stable
- Phasing out our old data volumes
- Created additional command line tools (shortcuts) to simplify some operations:
  - im: retrieves metadata for a file in the active tier based on its path in the Linux namespace
  - ngsactive.py: enables batch change of file active and restrict states
  - actreport.py: identifies files in the active tier that have not been recently accessed
- User roles and permission scheme needs to be refactored
  - Ingesters cannot activate/deactivate files if another ingester loaded them
  - Establish a data activator role distinct from data ingester role

Future Interests

- Upgrade iRODS version and refactor to leverage Python API
- Try the Audit Plugin
- Implement an archive data tier…perhaps with cloud storage and using the iRODS Storage Tiering Framework
- Explore Globus endpoint integration
- Process for adding analysis results data

Thanks

- Syngenta NGS Data Ingesters
  - Kirk Burthey
  - Bob Dietrich
  - Tonya Severson
- Syngenta HPC Team
  - Sean Korb
  - Marcelo Paparella
- RENCI
  - Jason Coposky
  - Cesar Garde
  - Terrell Russell
- Syngenta Informatics Platforms
  - Chris Martin
- Syngenta NGS Platform
  - Lucio Garcia
  - Laura Kavanaugh
  - Raja Kota
  - Ellis Merlin
  - Paul Travis