JGI JOINT GENOME INSTITUTE
UNITED STATES DEPARTMENT OF ENERGY

JAMO

JGI Archive and Metadata Organizer

MSST 2018

JGI Archive and Metadata Organizer (JAMO)

- JAMO's Beginnings
- What is JAMO?
- JGI Analysis Tracker
- What Files and Metadata are Stored
- System Stats

JAMO's Beginnings

- Fall 2012
  - Discussion on centralized system started
  - Visited WashU's Genome Center
- Winter 2012
  - Met with all science programs to discuss needs
  - Reviewed options (iRODS, BeSTMaN, etc)
- Spring 2012
  - Gap Analysis of needs unmet by existing SDM system
- Summer 2013
  - JAMO comes online
  - Raw sequence data moved into JAMO
  - JAMO used to facilitate retirement of file system
- Fall 2013
  - Start adding JGI groups to JAMO
  - JGI Analysis Tracker (JAT) Released
- Early 2014
  - 80% of JGI groups on JAMO

What is JAMO?

- **Managed repository of files and data**
- **Hierarchical metadata system**
- **Data provenance system**
- **Archive and retrieval system**
- **File lifecycle system**
- **Manages primary and secondary storage systems**
- **RESTful API metadata broker**
  - sharable, cascadable, user definable
- **Publish/Subscribe service**
- **Command line query tool and RESTful API**
- **User definable custom search and result/reporting system**
- **And more…**

JAT: JAMO Analysis Tracker

- **Layer on top of JAMO to manage and group files into work/analysis units**
- **Used interactively to:**
  - Find necessary inputs to perform analysis activities
  - Help users find/manage the Analysis Project/Task IDs
  - Helps users manage their disk footprint/quota
  - Helps users define necessary metadata and sources for tasks
  - Helps users define output file collection
  - Stages work for incomplete analysis
  - Provides backup services for data that needs to be purged
  - Loads datasets into JAMO on analysis completion
  - Automated email available on analysis completion
- **Template driven**
- **User contributed templates**
- **Metadata key and result file validation**
- **Continuous integration and validation for templates**
- **Manages data submissions into JAMO for users**
- **And more…**

What Files and Metadata are Stored?

- **DnA (Data ’n Archive) managed repository contains:**
  - Outputs from every group (final products)
  - Data that is shared across the organization
  - Data that is served by the portals
  - Data necessary to have reproducible processes/pipelines and science
- **Output files**
- **References to source data used to create output**
- **Software versions and parameters**
- **Portal virtual directory structure**
- **Group internal metadata**
- **JGI global identifiers**
- **Any other metadata needed to provide provenance and experiment repeatability**

JAMO Stats

- JAMO Data
  - 7.8 million files/tars in JAMO
  - 700+ million files collectively (counting tar contents)
  - 6.4 PB on tape (primary copy)

- Portal
  - Portal has 14k users
  - 500 users use GlobusOnline
  - 2 million files in JAMO visible on Portal
    - 1/4 of files in JAMO, 3.4 PB of data
  - 3,000+ distinct Portal users have downloaded data from JAMO over the last 12 months

- 1 million API calls per week made to JAMO