NASA Use Case

CISTO Experiences with an iRODS-Based Data Management Service

NASA Goddard Space Flight Center

Computational & Information Sciences & Technology Office (CISTO)

Glenn Tamkin

CISTO Mission

- *Traditional*
  - Enable scientists to increase their understanding of the Earth and the universe by providing state-of-the-art high performance computing, storage, network, and application solutions
  - Provide large-scale compute engines, analytics, data sharing, and high-end computing services
- *Future*
  - Develop a data services capability to better support the climate research communities and prepare the way for technology advances

Challenges

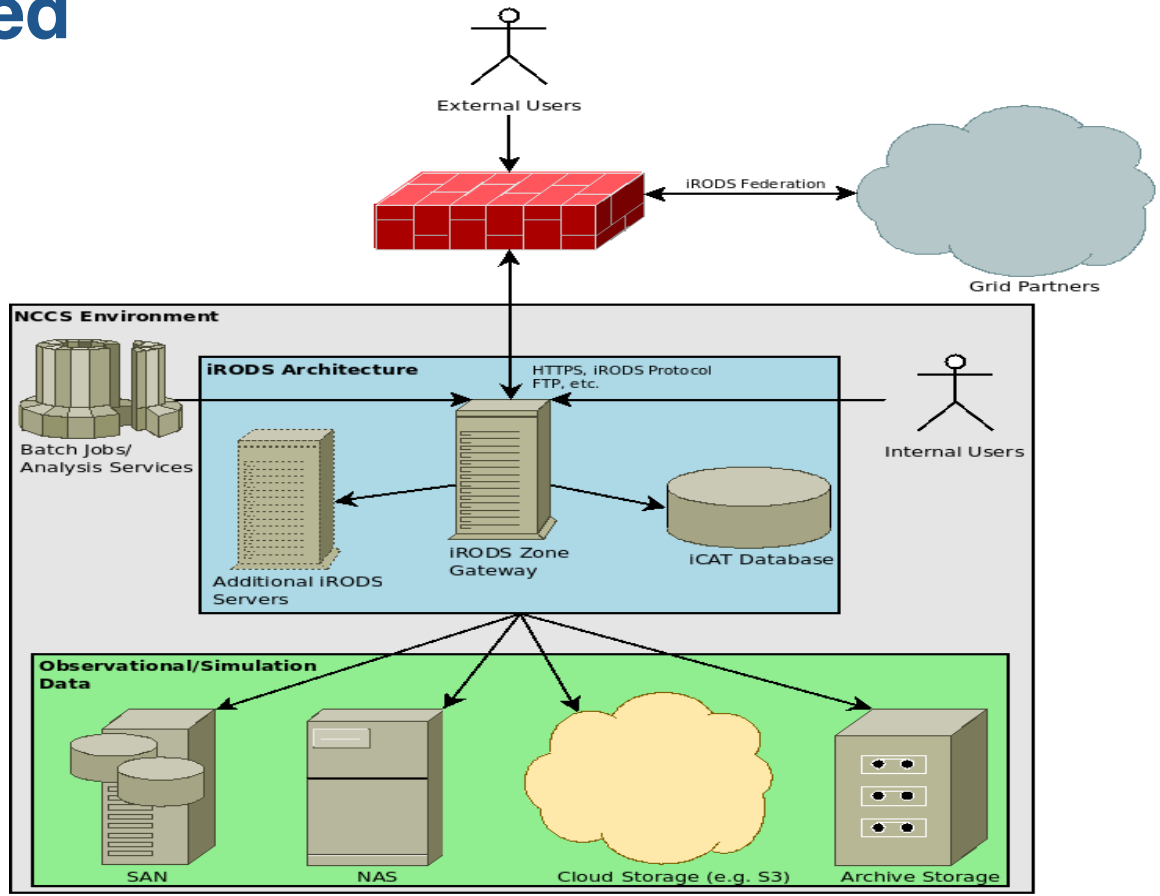
- *Finding* observational and model data for use in climate and weather studies
- *Accessing* the geographically distributed data
- *Managing* the massive digital holdings, which are measured in petabytes and hundreds of millions of files
- *Maintaining* the data, which must often be preserved for decades
- *Supporting* data sharing, data publication, and data stewardship

Technology - integrated Rule-Oriented Data System (iRODS)

- Open source data grid software developed by the Data Intensive Cyber Environments (DICE) group, University of North Carolina
- Targets large repositories and digital preservation
- Supports the federation of independent, distributed collections
- Supports server-side workflows that are implemented by chaining execution rules together based on data policies
- Includes features such as domain-specific validation, automatic replication, and digital signature/checksum computation
- Validates assertions about data such as integrity and authenticity

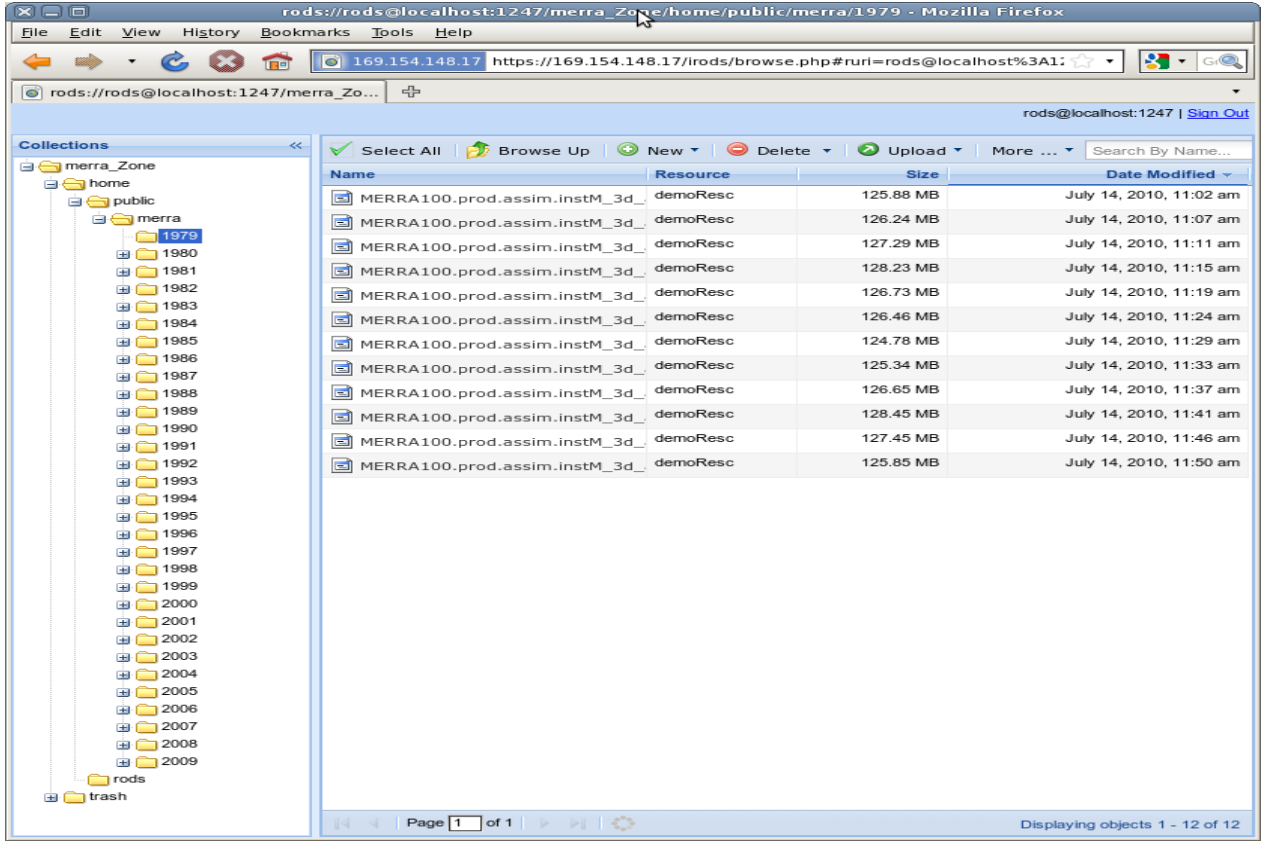
CISTO Test Bed

- iRODS abstracts physical location of data
- iRODS assists with archive management



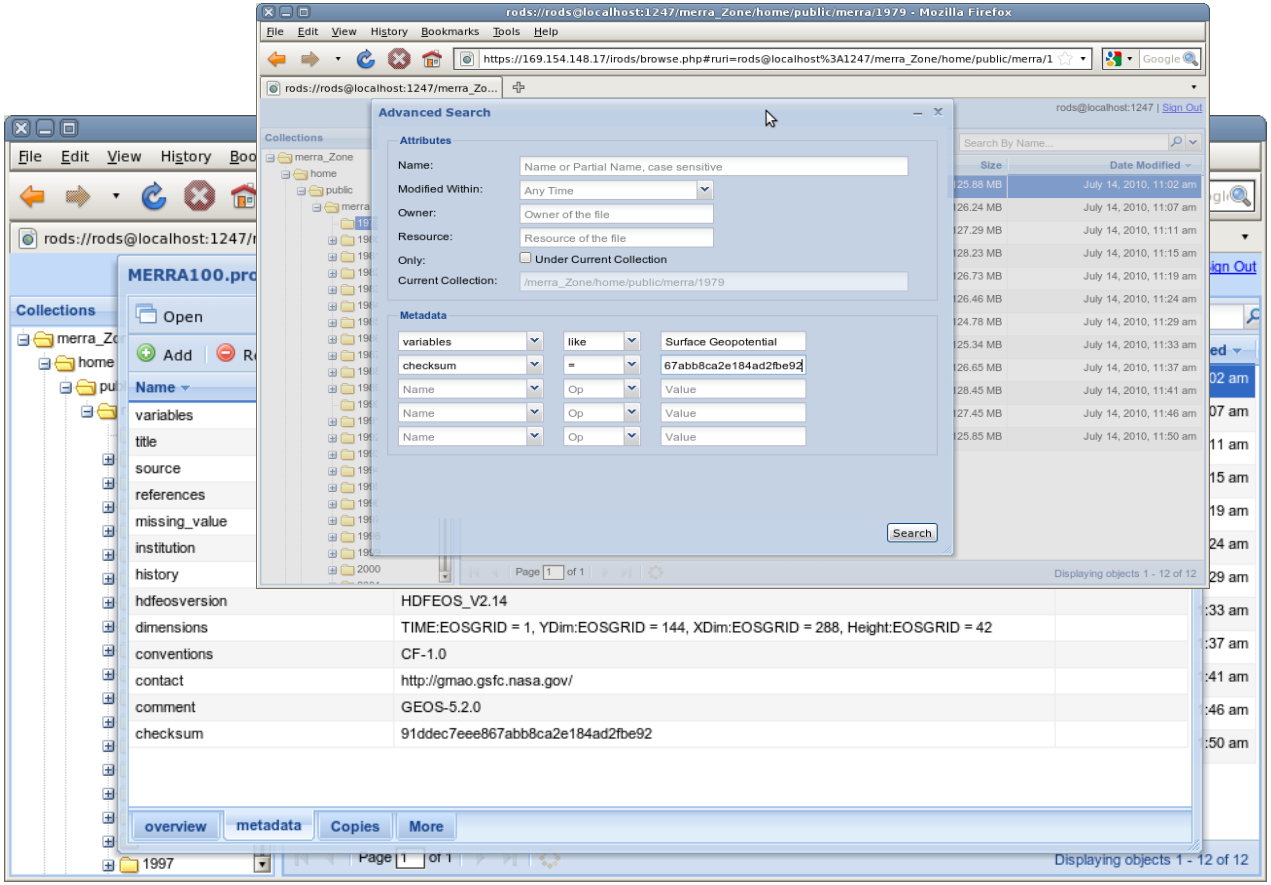
Ingest/Registration

- iRODS rules and microservices allow data to be stored in configurable collections based on data policies
- Replication to backup storage resources also supported



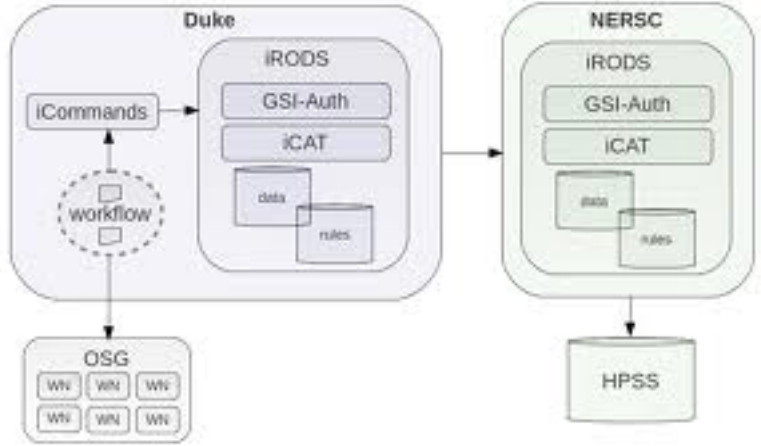
Search

- iRODS rules and microservices can be used to assign metadata
- iRODS provides advanced search capabilities over the metadata



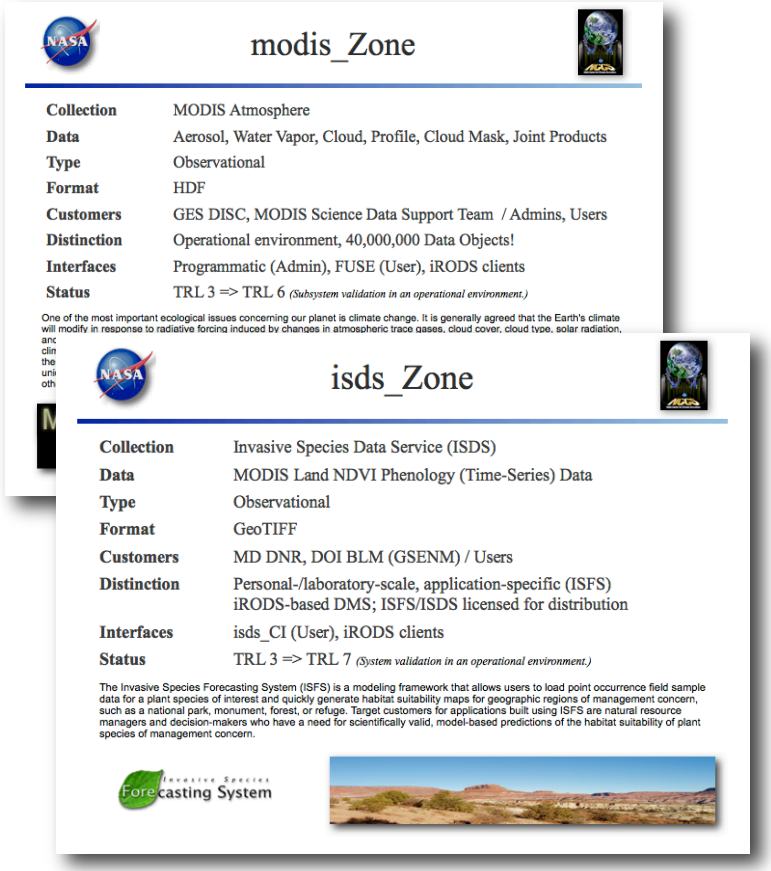
Federation

- Federation allows separate iRODS Zones (iRODS instances) to become integrated.
- Each zone is a separate iRODS instance, administered separately, but the users in the multiple zones, if given permission, will be able to access data and metadata in the other zones.
- No user passwords are exchanged, as each system will, in a secure manner, check with the user's local zone for authentication when the user connects.



Preliminary Tests – Observational Data

- Developed an iRODS data grid that publishes Moderate Resolution Imaging Spectro-radiometer (MODIS) observational data
  - 54 million registered files, 630 TB of data, and over 300 million defined metadata values
- Developed an iRODS data grid that focuses on a small-scale, multi-product, application-specific data service
  - The Invasive Species Data Service (ISDS) manages a collection of MODIS data products for ecological forecasting applications



modis_Zone

Collection	MODIS Atmosphere
Data	Aerosol, Water Vapor, Cloud, Profile, Cloud Mask, Joint Products
Type	Observational
Format	HDF
Customers	GES DISC, MODIS Science Data Support Team / Admins, Users
Distinction	Operational environment, 40,000,000 Data Objects!
Interfaces	Programmatic (Admin), FUSE (User), iRODS clients
Status	TRL 3 => TRL 6 *(Subsystem validation in an operational environment.)*

One of the most important ecological issues concerning our planet is climate change. It is generally agreed that the Earth's climate will modify in response to radiative forcing induced by changes in atmospheric trace gases, cloud cover, cloud type, solar radiation,

isds_Zone

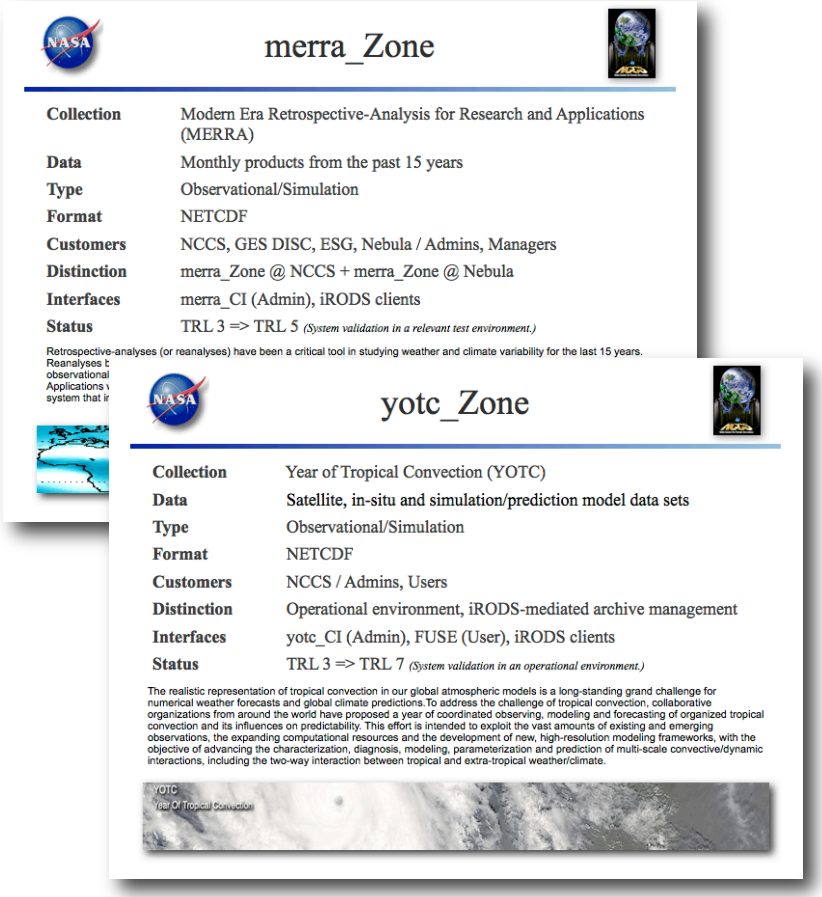
Collection	Invasive Species Data Service (ISDS)
Data	MODIS Land NDVI Phenology (Time-Series) Data
Type	Observational
Format	GeoTIFF
Customers	MD DNR, DOI BLM (GSENM) / Users
Distinction	Personal-/laboratory-scale, application-specific (ISFS) iRODS-based DMS; ISFS/ISDS licensed for distribution
Interfaces	isds_CI (User), iRODS clients
Status	TRL 3 => TRL 7 *(System validation in an operational environment.)*

The Invasive Species Forecasting System (ISFS) is a modeling framework that allows users to load point occurrence field sample data for a plant species of interest and quickly generate habitat suitability maps for geographic regions of management concern, such as a national park, monument, forest, or refuge. Target customers for applications built using ISFS are natural resource managers and decision-makers who have a need for scientifically valid, model-based predictions of the habitat suitability of plant species of management concern.

Invasive Species Forecasting System

Preliminary Tests – Simulation Data

- Developed an iRODS data grid that manages Modern Era Retrospective-Analysis for Research and Applications (MERRA) simulation data
  - 360 files, 47 GB of data, and 4000 metadata values
- Developed an iRODS data grid that publishes Year of Tropical Convection (YOTC) data sets
  - 134,000 files, 12 TB of data, and 400,000 metadata values



merra_Zone

Collection	Modern Era Retrospective-Analysis for Research and Applications (MERRA)
Data	Monthly products from the past 15 years
Type	Observational/Simulation
Format	NETCDF
Customers	NCCS, GES DISC, ESG, Nebula / Admins, Managers
Distinction	merra_Zone @ NCCS + merra_Zone @ Nebula
Interfaces	merra_CI (Admin), iRODS clients
Status	TRL 3 => TRL 5 *(System validation in a relevant test environment.)*

Retrospective-analyses (or reanalyses) have been a critical tool in studying weather and climate variability for the last 15 years.

yotc_Zone

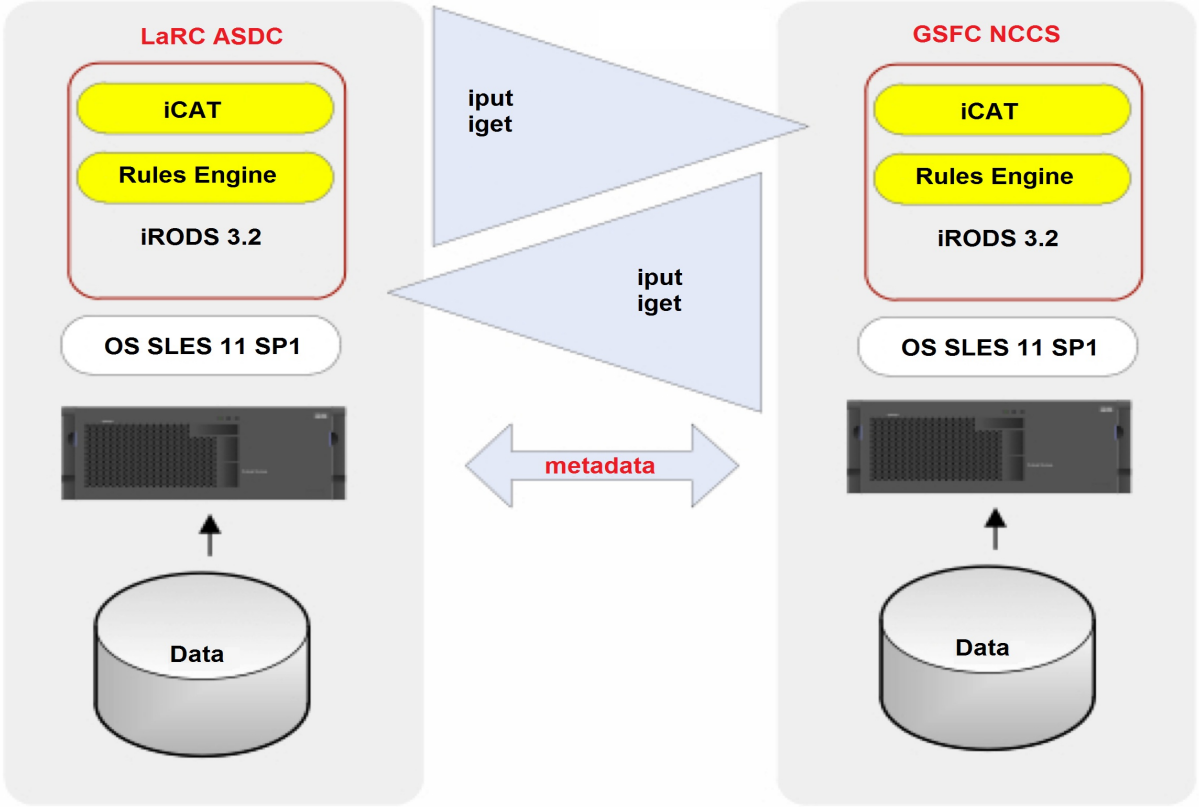
Collection	Year of Tropical Convection (YOTC)
Data	Satellite, in-situ and simulation/prediction model data sets
Type	Observational/Simulation
Format	NETCDF
Customers	NCCS / Admins, Users
Distinction	Operational environment, iRODS-mediated archive management
Interfaces	yotc_CI (Admin), FUSE (User), iRODS clients
Status	TRL 3 => TRL 7 *(System validation in an operational environment.)*

The realistic representation of tropical convection in our global atmospheric models is a long-standing grand challenge for numerical weather forecasts and global climate predictions.To address the challenge of tropical convection, collaborative organizations from around the world have proposed a year of coordinated observing, modeling and forecasting of organized tropical convection and its influences on predictability. This effort is intended to exploit the vast amounts of existing and emerging observations, the expanding computational resources and the development of new, high-resolution modeling frameworks, with the objective of advancing the characterization, diagnosis, modeling, parameterization and prediction of multi-scale convective/dynamic interactions, including the two-way interaction between tropical and extra-tropical weather/climate.

NASA LARC/GSFC iRODS Federation

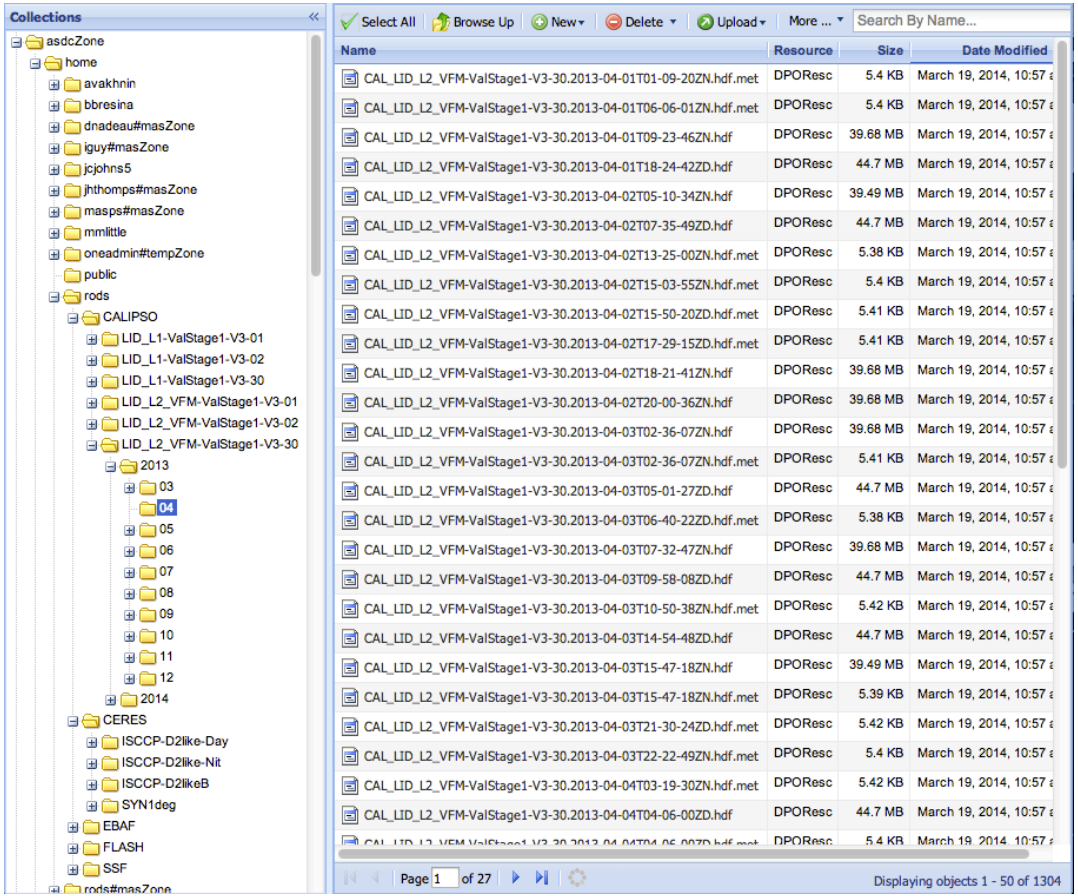
- Stakeholders
  - NASA's Atmospheric Science Data Center (ASDC) and Center for Climate Simulation (NCCS) both ingest, archive, and distribute data that is essential to stakeholders including the climate research community, science applications community, and a growing community of government and private-sector customers who have a need for atmospheric and climatic data.
- Goals
  - To implement a data federation ability to improve and automate the discovery of heterogeneous data, decrease data transfer latency, and meet customizable criteria based on data content, data quality, metadata, and production. Another goal of the federation is to support/enable applications and customers that require the integration of multiple heterogeneous data collections.

NASA LARC/GSFC Federation Architecture



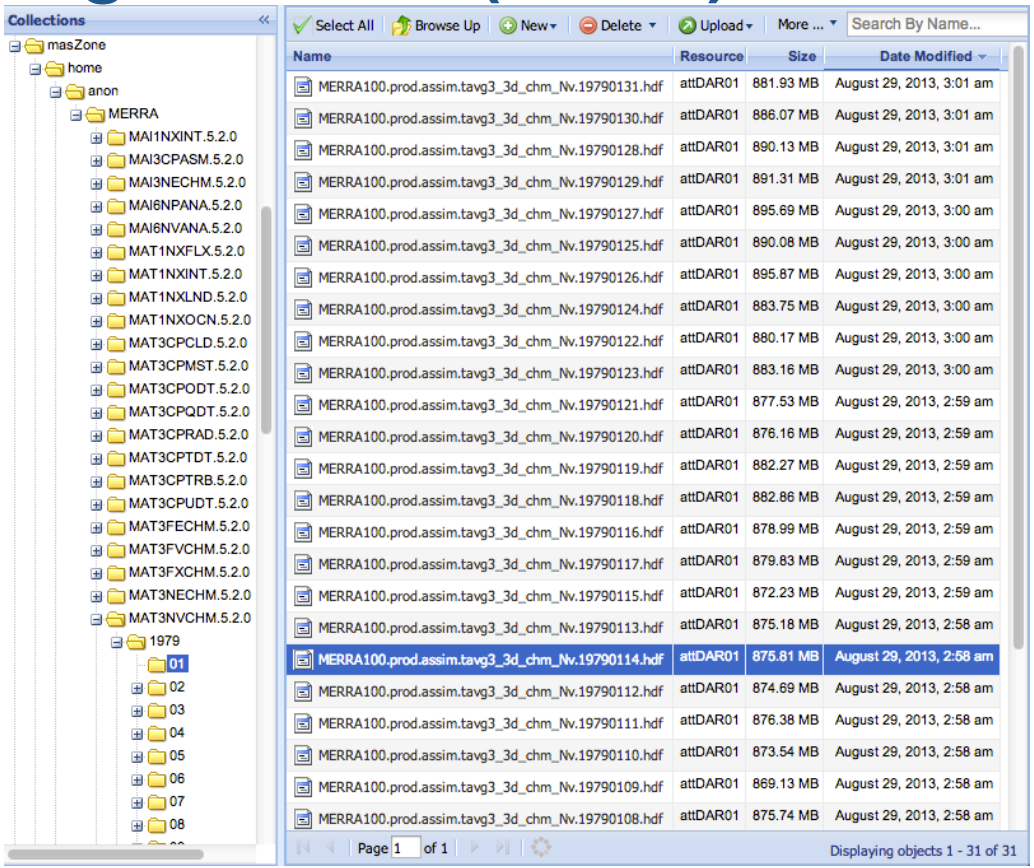
Langley Research Center (LARC) Zone

- The LARC Atmospheric Science Data Center (ASDC) Zone contains varies scientific collections (e.g., CALIPSO, CERES)
- ASDC's Data Products Online (DPO) is federated with the GSFC/MAS Zone.



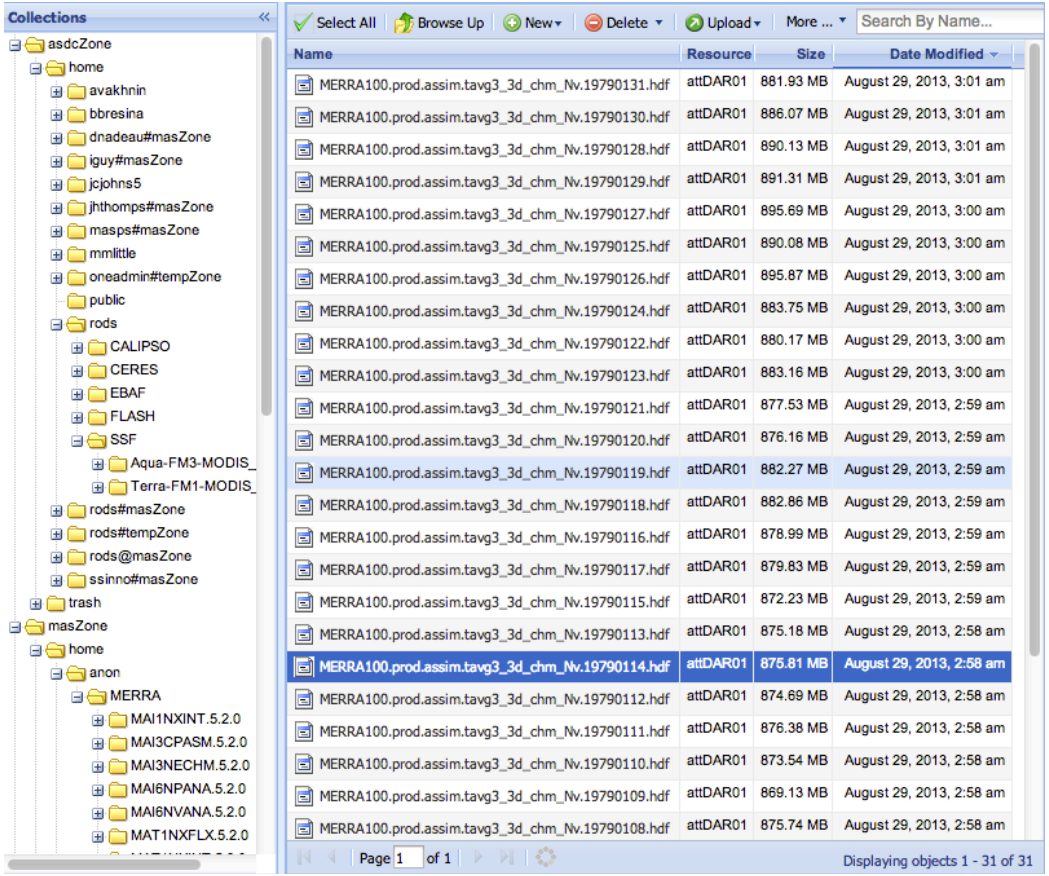
Goddard Space Flight Center (GSFC) Zone

- The GSFC MERRA Analytic Service (MAS) Zone contains the full MERRA collection
- All NetCDF metadata is registered for each file
- Modern-Era Retrospective Analysis for Research and Application (MERRA) is federated with the LARC/ ASDC Zone



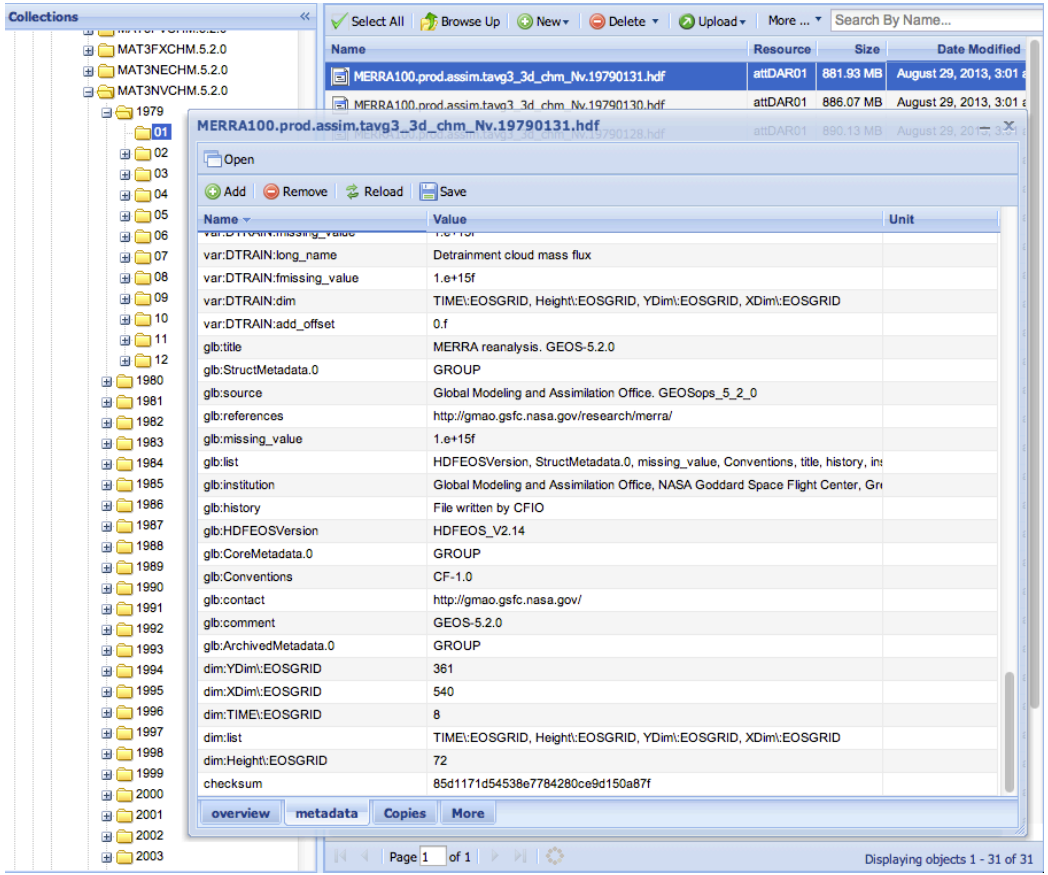
LARC/GSFC Federation

- Both iRODS Zones are accessible together
- Users can access the data via web browser, command line, FUSE file system, etc.



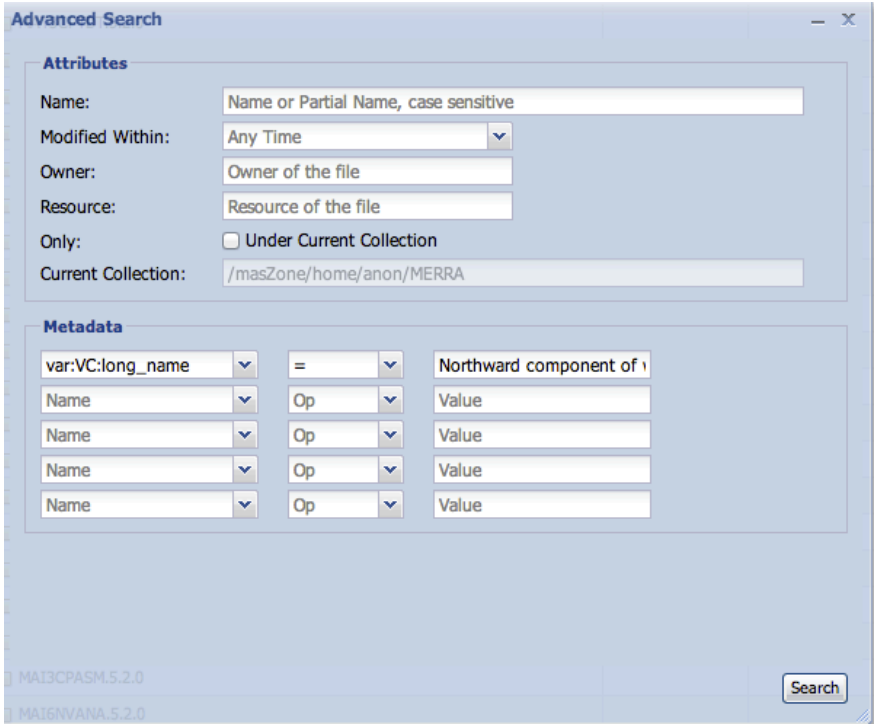
LARC/GSFC Metadata

- All NetCDF files are harvested for metadata at registration time using iRODS microservices and rules
- Users can view and search across the metadata
- Custom/derived metadata can be applied to any collection or file



LARC/GSFC Search

- Files can be searched by attribute or metadata



Summary

- iRODS federation can be a powerful mechanism for creating ensembles of distributed data
- Collections are still managed by the data owners, but iRODS provides mechanisms for creating and enforcing policies that control external access within the federation
- iRODS abstracts the advertised collection and the physical location of the data. As a result, the data can be relocated or extended without disturbing the federated consumer. In fact the consumer needn't be aware of the federation, as it appears as one large ensemble
- This approach enables parallel downloads of datasets from selected replica servers that can be geographically dispersed, but still accessible by users worldwide

Next Steps

- Inclusion of a new member to the federation
- Extended access to files on the ASDC's tape archive
- Collection expansion to include LandSAT and MODIS data
- Utilization of the NetCDF plugin to assign metadata to arbitrary NetCDF/HDF4/HDF5 data.
- Testing/utilization of the iRODS HDFS plugin (Hadoop access)
- Testing/utilization of methods to register remote data in the Public Domain via iRODS MicroService Objects
- Automated file registration using operating system level triggers (i.e., 'iNotify')

Future ASDC iRODS Configuration