

Using Corral for Research Data Management

Chris Jordan

October 25, 2012

Outline

- Introduction to Corral
- Corral/TACC Data Management Services
- Gaining access to Corral/Appropriate uses
- Using the file system
- Using iRODS
- Access Control and Security
- Metadata and Data Management

Corral/UT Data Repository

- “Corral” encompasses a collection of storage systems and services for data management
- UT Data Repository - the newest and largest component of Corral
- 5PB online storage in Austin and Arlington
- Connection to new 10Gb Research Network (in progress on many campuses)
- Multiple access mechanisms and services

Corral Speeds and Feeds

- Over 4,000 hard drives in all
- Dell servers, DDN disk systems, IBM GPFS
- 8 nodes managing file systems, 10 nodes managing data transfer and services
- 20 Gigabyte per second internal performance
- 20 Gigabit per second network connections
- Multiple layers of redundancy and replication
- Full Geographical Replication coming soon

Notes on Performance

- Corral has very high potential performance
- Mostly, local network is the limiting factor
- Different tools have different performance
- You will likely see between 1MB/s & 100MB/s
- UT Research Network will upgrade campuses to 10Gb (1 Gigabyte/sec)
- If performance to Corral is a major concern, contact your local network administrator

Corral Intended Use

- Corral is a Research Data Repository
- Primarily intended for storage of data used in (your own or others) research
- Not intended for document collaboration
- Not intended for administrative data
- Not a backup service
- Sharing data with others is encouraged
- Please contact us with questions

Corral and HPC

- Corral file systems are directly accessible from Lonestar
- Lonestar has local file systems for HPC
- Corral is appropriate for long-lived and larger data collections used in HPC
- Also, shared data collections used by multiple groups

Corral Major Services

- “Simple File System”
 - Looks like any desktop or server file system
 - Can be directly accessed from Lonestar
 - Group and user permissions, Access Control Lists
- iRODS Data Management System
 - Provides a “virtual” FS over multiple resources
 - Accessed through special client tools
 - Supports extended metadata and policies

Corral – Additional Services

- Database hosting
 - Postgres and MySQL
- Web Hosting
 - “Open” collections (no restrictions)
 - Semi-public collections (large projects, etc)
- Many other complex services can be provided, but require special arrangements

Web Access Modes

- Most Web access modes require iRODS
- Private Web Access – All iRODS users get access to web browser interfaces
- Public Web Access – iRODS users may request a public web area
 - /corralZ/web/ProjectName
 - All data stored in this collection becomes web-accessible via a URL

Data Management Services

- All research involving digital data should involve data management planning
- NSF/NIH grants now require such planning
- Metadata, digital preservation, data curation may all be necessary regardless of grant requirements
- Corral is one of a number of potential resources

Consulting with TACC

- Basic support of Corral is free
- Additional scenarios:
 - Collaborations on grants with a significant technology component
 - Consulting for projects with complex metadata and/or data curation issues
 - Assistance with digital preservation projects
- Contact us directly with questions

Getting Access to Corral

- UT System PIs (i.e. researchers at any UT institution) can have allocations up to 5TB on request, with annual renewal
- Larger allocations and long-term commitments: \$250/TB/year
- All user setup and allocations are processed through the TACC user portal

TACC User Portal

The screenshot displays the TACC User Portal interface. At the top, the browser address bar shows the URL: https://portal.tacc.utexas.edu/group/tup/projects-and-allocations?p_id=taccallocations_WAR_taccallocationsportlet_INSTANCE_g8C&p_p_lifecycle=0&p_p_st. The page title is "Projects and Allocations - TACC User Portal".

The main header features the "TACC USER PORTAL" logo on the left and a "Sign Out" link on the right. Below the logo is a navigation menu with the following items: HOME, NEWS, RESOURCES, ALLOCATIONS, DOCUMENTATION, TRAINING, CONSULTING, and ABOUT. The "ALLOCATIONS" menu item is currently selected, and a dropdown menu is visible with options for "PROJECTS AND ALLOCATIONS" and "ALLOCATIONS OVERVIEW".

Below the navigation menu, there is a "TACC Projects and Allocations" section. It includes a "create new project" button and a "Need help?" link. The main content area displays a list of project categories, each with a list of sub-items:

- TG-STA110007S (PI)**
 - Project Information
 - Grants
 - Publications
 - Allocations
 - Users
 - Usage
- A-stockade (Delegate)**
 - Project Information
 - Grants
 - Publications
 - Allocations
 - Users
 - Usage
- 20120110UTRC (Standard)**
 - Project Information
 - Grants
 - Publications
 - Allocations
 - Users
 - Usage
- 20120217UTRC (Standard)**
 - Project Information

The browser address bar at the bottom of the screenshot shows the URL: <https://portal.tacc.utexas.edu/group/tup/allocations>.

The TACC Allocations Process

- Sign up for an account
- Create a project in the user portal
- Request allocations on Corral and any other resources
- TACC staff review each request for:
 - Sufficient information/appropriate usage
 - Data Security issues
 - Research Justification (NOT a peer review)

Getting an Allocation

- Upon approval, you will receive an e-mail with appropriate additional information
- You may receive follow-up questions about specific services or setup issues
- Questions or concerns? Contact us early
- Allocations and setup process usually takes a few days
- So, now you have an allocation ...

Using the Corral file systems

- For UNIX users, this will be very familiar
- Project path, looks like:
 - /corral-repl/utexas/MyProject
- Can be accessed from Lonestar or from Corral login node
 - login1.corral.tacc.utexas.edu
- Replication is handled transparently

Corral and TACC File Systems

- “Corral” consists of multiple parallel file systems
- These are accessible from various systems
- Each system (including Corral login) has its own HOME area
- Don’t store your data there!
- Lonestar, other HPC platforms have dedicated high-performance file systems

File System Characteristics

- Simple, fast, familiar to UNIX users
- Easy access from HPC and other systems
- Not great for sharing of data outside of TACC
- Quotas used to enforce allocation limits
- Two options :
 - /corral-repl for replicated storage
 - /corral-tacc for TACC-only storage (may be faster)

SSH and SCP

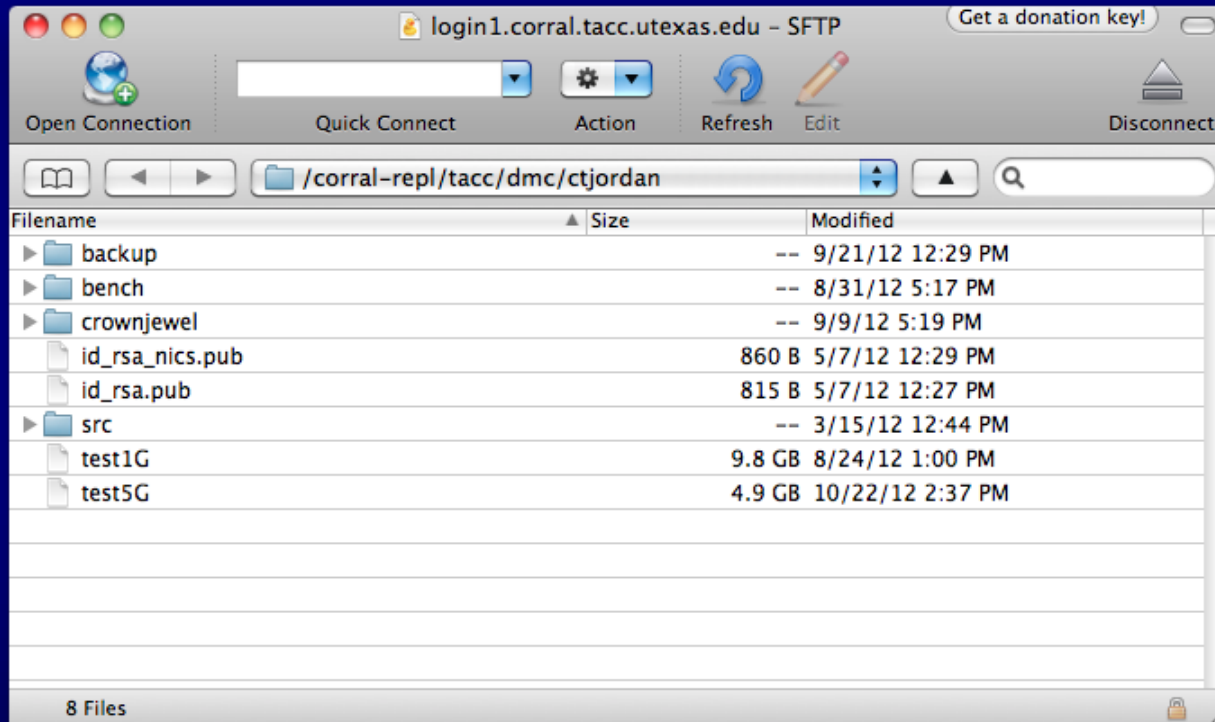
- All TACC Systems are accessed via Secure Shell (SSH)
- Mac and Linux users – congratulations, you already have SSH
- Windows – PuTTY, other free clients
- Cyberduck for file transfer (free, graphical, work on Mac and Windows)

Cyberduck SSH Login

The image shows a screenshot of the Cyberduck SFTP (SSH File Transfer Protocol) login dialog box. The dialog is titled "SFTP (SSH File Transfer Protocol)". It contains the following fields and options:

- Server:** login1.corral.tacc.utexas.edu
- Port:** 22
- URL:** sftp://ctjordan@login1.corral.tacc.utexas.edu:22/
- Username:** ctjordan
- Password:** [masked with dots]
- Anonymous Login
- Add to Keychain
- Buttons:** ? (help), Cancel, Connect
- More Options:**
 - Path:** /corral-repl/utexas/myproject
 - Connect Mode:** Default
 - Encoding:** Default
 - Use Public Key Authentication
 - No private key selected

Cyberduck SSH Window



Using iRODS

- Integrated Rule-Oriented Data System
- Creates “virtual directory” on top of existing file systems
 - i.e. uses file systems discussed in previous section
- Provides enhanced metadata, policy, and web access support
- APIs for direct access from applications

When to use iRODS

- No need for direct access from Lonestar
- No need for UNIX-style access
- Need for metadata/search support
- Need for web-based/special client access
- Examples:
 - Digital Preservation of Cultural Objects
 - Natural History Collection Images

How iRODS is used

- No inherent “file system” interface
- Hierarchical structure of “collections”
- “i-Commands” for manipulating data
 - Must have ~/.irods/.irodsEnv file
- Special commands for metadata creation, query functions, permissions
- Web/WebDAV interface and GUI clients
- Java, other APIs

Getting iRODS clients

- iRODS home page <http://www.irods.org>
- Unix clients are straightforward to build
- Binaries are available for many platforms
- Also, other clients including iDrop
- iDrop is Java-based, multi-platform
- APIs can also be found here for integration into your own code (C, Python, Java, etc)

i-Command examples - 1

- “`iput -v -R gpfs-repl -r -K <directory>`”
 - “-v” Verbose
 - “-R <resource> to select a specific resource
 - “-r” Recursive
 - “-K” create and verify checksums”
- “iget” command uses same options to retrieve data

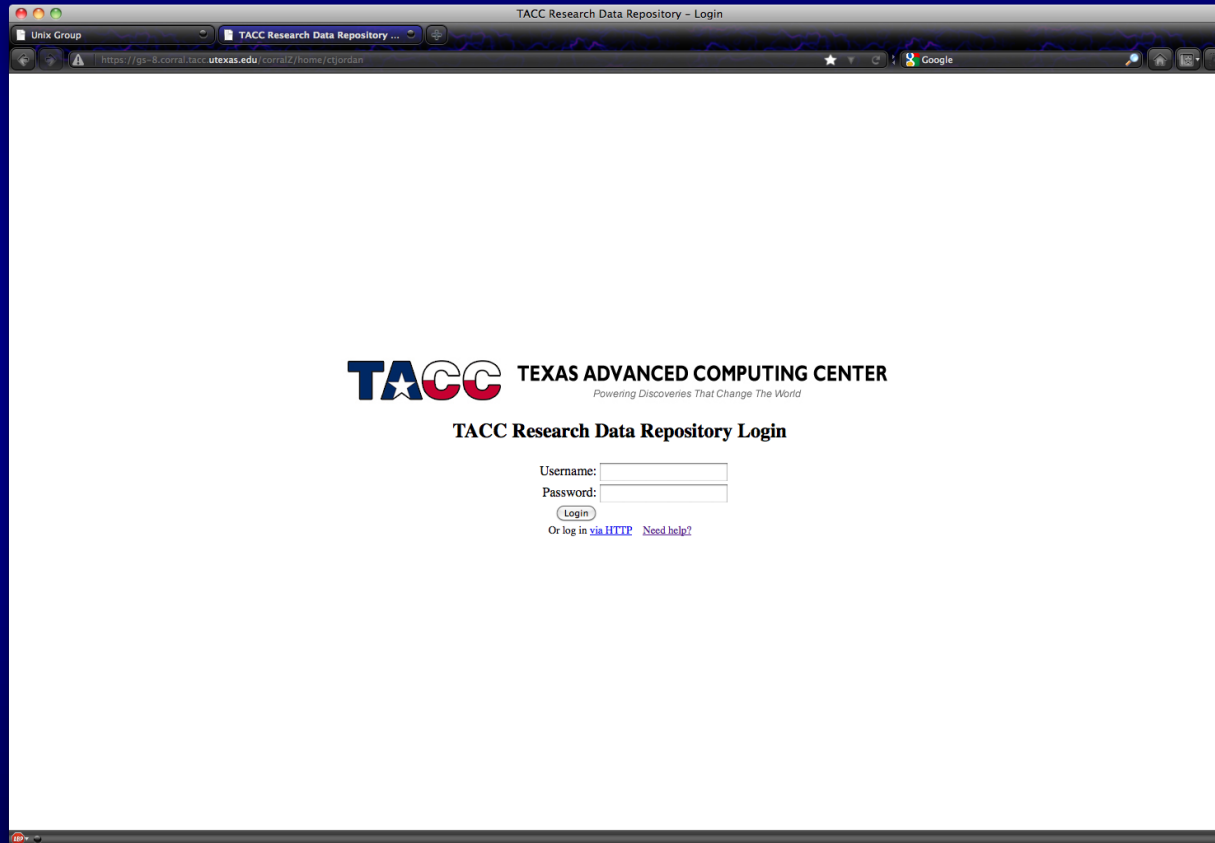
i-Command examples - 2

- “ichmod own MyGroup <collection>”
 - Gives full permissions to all members of the group
 - “ichmod inherit <collection>” makes all new files and folders inherit these permissions
- iquest “select COLL_NAME where DATA_NAME like ‘MyFile%’”
 - Gives the location of all files beginning with MyFile
 - Many other SQL-Like parameters
 - Best used in scripts built around functions

i-Command examples - 3

- irsync
 - Synchronizes entire directory tree into iRODS
- imeta
 - Creates, manipulates and locates metadata
- irepl
 - Creates replicas of existing data within iRODS
- ilsresc
 - List all available storage resources

iRODS Web Interface Login



The screenshot shows a web browser window with the title "TACC Research Data Repository - Login". The address bar displays the URL "https://gs-8.corral.tacc.utexas.edu/corralZ/home/cjordan". The page content includes the TACC logo (Texas Advanced Computing Center) with the tagline "Powering Discoveries That Change The World". Below the logo is the heading "TACC Research Data Repository Login". The login form consists of two input fields: "Username:" and "Password:". A "Login" button is positioned below the password field. At the bottom of the form, there is a link "Or log in via HTTP" and a link "Need help?".

TACC TEXAS ADVANCED COMPUTING CENTER
Powering Discoveries That Change The World

TACC Research Data Repository Login

Username:

Password:

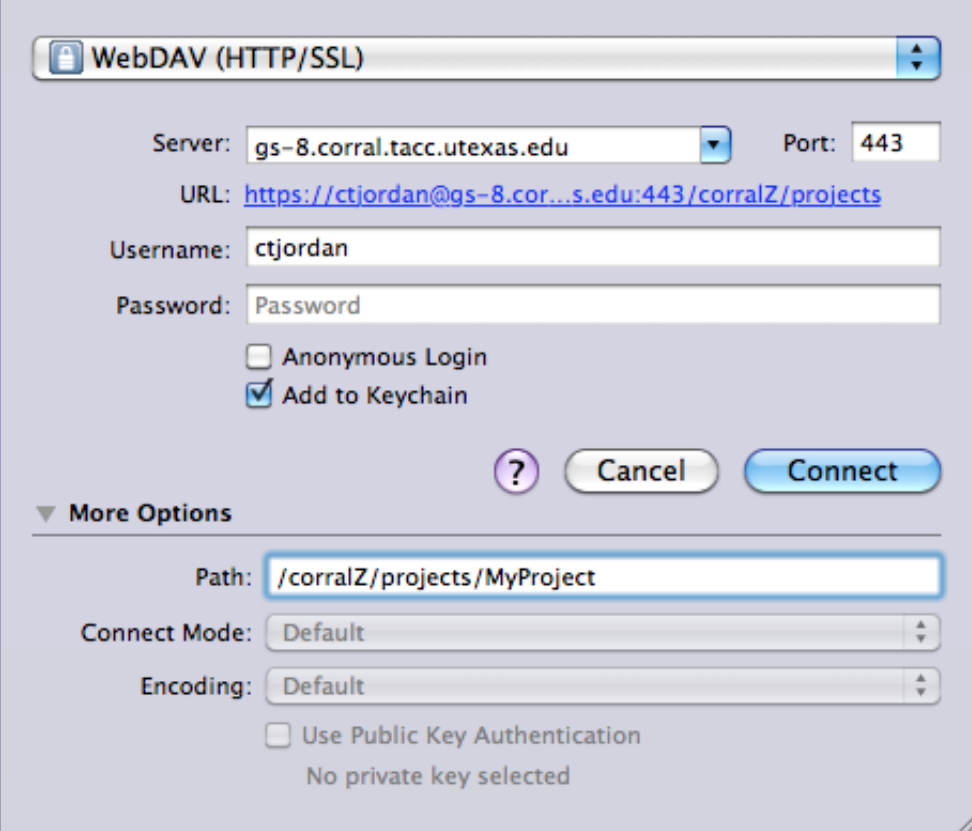
Or log in via [HTTP](#) [Need help?](#)

iRODS Web Interface

The screenshot shows a web browser window with the URL `https://gs-8.corral.tacc.utexas.edu/corralZ/home/cjordan#1351024046513`. The page title is "TACC Research Data Repository - Start Page - Texas Advanced Computing Center". The user is logged in as "cjordan". The breadcrumb path is `/corralZ > home > cjordan > backups`. There are 5 items listed, 0 items selected. The interface includes buttons for "Upload File" and "Create Directory". A table lists the contents of the directory:

Name	Last Modified	Size	QuickShare	Select all
... Parent Directory				
arctos	Fri, 28 Sep 2012 19:14:24 GMT			Access Control
db	Fri, 28 Sep 2012 16:52:05 GMT			Metadata
desktop	Fri, 28 Sep 2012 17:01:40 GMT			Replicas
irods	Fri, 28 Sep 2012 19:01:55 GMT			Rename
www	Sun, 30 Sep 2012 13:20:52 GMT			Copy
				Move
				Delete

iRODS Cyberduck Configuration



The image shows a configuration window for WebDAV (HTTP/SSL) in Cyberduck. The window title is "WebDAV (HTTP/SSL)". It contains the following fields and options:

- Server: gs-8.corral.tacc.utexas.edu
- Port: 443
- URL: <https://ctjordan@gs-8.cor...s.edu:443/corralZ/projects>
- Username: ctjordan
- Password: Password
- Anonymous Login
- Add to Keychain
- Buttons: ? (help), Cancel, Connect
- More Options section:
 - Path: /corralZ/projects/MyProject
 - Connect Mode: Default
 - Encoding: Default
 - Use Public Key Authentication
 - No private key selected

iRODS “Zones”

- iRODS has the ability to “federate” data grids
- Allows easy data transfer and replication between systems
- iPlant, XSEDE and many other large projects have their own such iRODS data grids
- Already support federation with some existing grids
- We can support new zones in some cases

Access Control and Security

- Corral supports both public and private data
- Can support private data NOT subject to special legislation
- For HIPAA/FERPA/etc data, please contact us to discuss special arrangements
- For all users, important to carefully manage security and access to data

File Permissions Basics

- Basic model: User/Owner, Group and World
- Group is equivalent to project
 - Users on multiple projects must be aware of this
- Read, Write, Execute
- Access Control Lists
 - Special mechanism to allow specific access to specific users
 - iRODS provides simpler management of ACLs

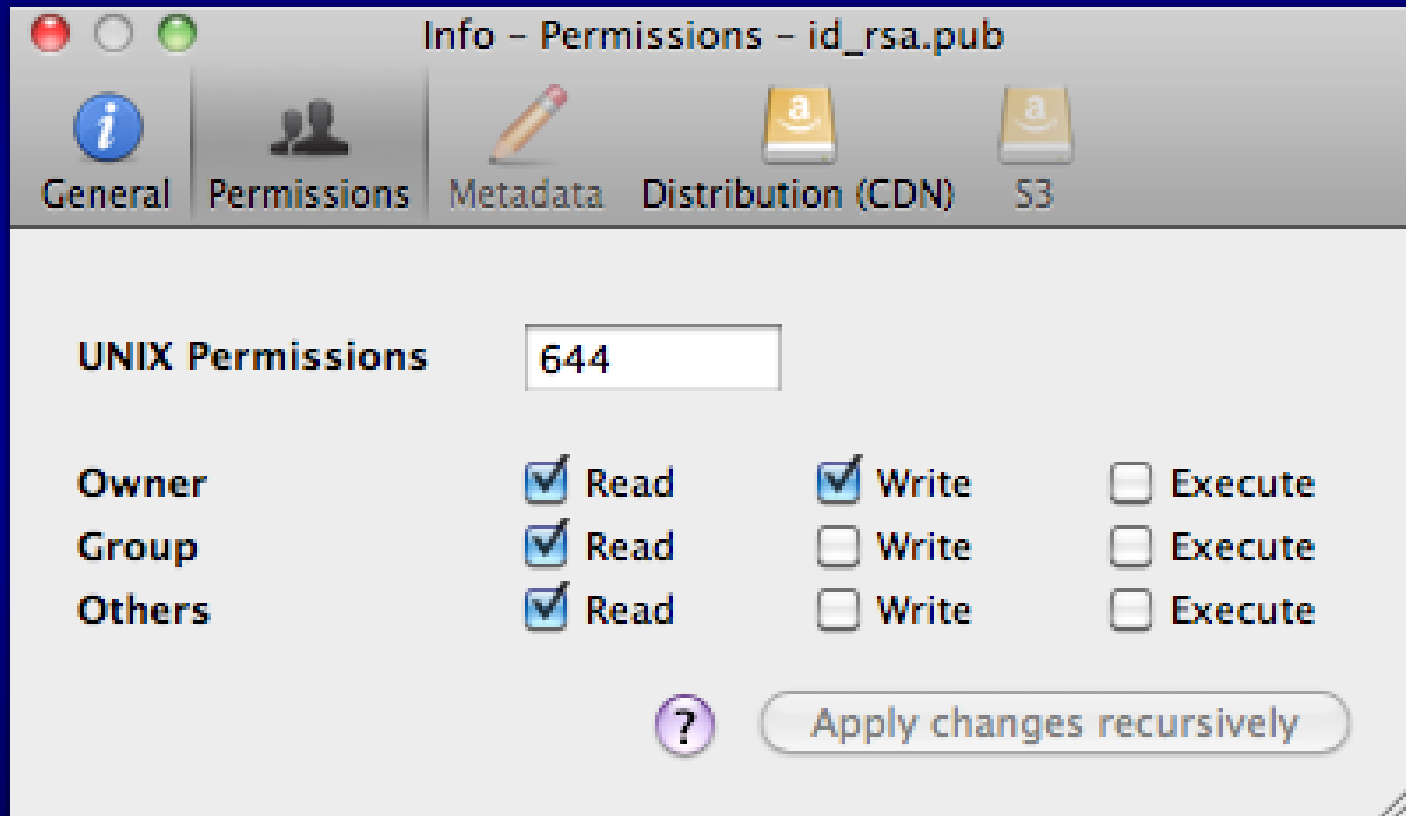
File Permission setup

- When you first get your allocation, your folder will (usually) have group and user access
- Should change this immediately if desired
- Unix: “chmod”
- iRODS: “ichmod”
- Cyberduck: “Info -> Permissions”

Basic chmod usage

- Simple form – what group you are referring to, add or subtract, which permissions
- “chmod g-rwx <folder>”
 - Removes all permissions for the group
- iRODS: “ichmod read <group> <folder>”
 - Grants read access to the group

Cyberduck Permissions Example



Ownership and Quotas

- Disk usage is determined by the file “owner”
 - Also controls access to data
 - Only an issue for users in multiple groups
- To ensure proper group ownership
 - “chgrp -R <mygroup> <myfolder>”
- Checking usage and quotas can be done through the TACC user portal (soon)

Metadata and Data Management

- “Data Management” simply implies thinking about your data ahead of time:
 - What will be created?
 - How will it be organized?
 - Who will access it?
 - How will they find it?
 - How will they retrieve it?
 - How do I ensure it’s still there when I need it?

Basic Data Management

- Folder hierarchy
 - What categories of data will you have?
 - What categories are meaningful when locating?
- Differing value/audience of data
 - Temporary, permanent, reference
 - Consider storing in different locations or using metadata to indicate retention period
 - Private/public data

Advanced Data Management

- Metadata, Curation, Workflows
- Workflow as crucial first step
 - What actions are taken on all or most data?
 - Can these be automated?
 - What metadata applies to all this data?
 - What are the different audiences for this data?
 - What can be thrown away/what must be saved?
 - Where in the research process are these events?

Data Management Plans

- Templates are available but may not address project-specific concerns
- Consider using reference repositories
- Corral can be an important part of workflows
- Web-based open distribution
- To ensure access and demonstrate seriousness, consider budgeting for storage
- We can help?

What is Metadata

- Metadata = Data about Data
 - Technical: File type, size, date created
 - Format: Encoding method, dimensions, length
 - Legal: Owner/Creator, License restrictions, Embargo periods
 - Provenance: Instrument/Sensor, Experimental conditions, sample preparation
 - Preservation: Checksums, Replica counts, etc
 - And many more

Creating Metadata

- Metadata can be created manually, but must have constraints
- Could be many ways of describing a file format, a sensor, a sample
- Automatic extraction from files
 - Image metadata created by digital cameras
- Constrained manual creation
 - Web forms, scripts, etc

Q&A



Chris Jordan

ctjordan@tacc.utexas.edu

data@tacc.utexas.edu

For more information:

www.tacc.utexas.edu

