TEXAS ADVANCED COMPUTING CENTER

DATA SERVICES AND NATIONAL CYBERINFRASTRUCTURE

Chris Jordan

TACC Data Management and Collections Group

January 24 2020

OUTLINE

- What is this TACC place anyway?
- Corral – A Unique Data Services Platform
- National Cyberinfrastructure Context
- Collaboration Models and Examples

TEXAS ADVANCED COMPUTING CENTER

- Organized Research Unit of the University of Texas at Austin
- Nowadays known for Frontera and Stampede (1 & 2) – Among the fastest and most-utilized high performance compute resources globally
- Diverse compute, visualization and data resources serving wide array of research needs, at all scales
- Not really about hardware – TACC is an institution with deep and broad expertise in advanced computing for research

TACC AND THE DATA SERVICES CONUNDRUM

- Need both data services (access) and storage (big magnets/silicon)
- (Almost) No funding for persistent, access-oriented storage
- Closest approach was Wrangler, a data-intensive compute platform with 8PB of "project-term" storage and a 4 year lifespan
  - No funding or alternative resource for data on Wrangler at end-of-life
- Tape archives at TACC and similar centers for large-scale, long-term, minimal-access data storage
  - Funding story is complicated at best
  - But these are success stories for keeping data– stretching back to the 1980s
- Corral – can we build and maintain storage infrastructure with stable financing?

CORRAL – THE STORY SO FAR

- Commissioned in 2009 as a ~1PB Lustre resource using private donation
  - Major advantages: no explicit lifespan for data, freedom to allocate based on research needs/partnerships, explicit tie to data services for access
- "Corral 2" four years later : 4PB GPFS with offsite replication
  - Funding from University of Texas System, served all 15 UT schools
  - Incorporated fee-based model for sustainability
- Corral 3 in 2016 expands to 12PB
  - At this stage, "Corral" is more a collection of data services/storage than a single system

CORRAL PRESENT AND FUTURE

- Now an established resource with a robust service model and a large user base
- Roughly 4-5 year refresh cycles for hardware
- Over 200 collections from PI-level to national scale (some over 10 years old)
- Trying to create something with a near-permanent lifespan, without claiming "forever" storage
- But "we" need more than one of these, for redundancy and robustness nationally
  - Still interested in partnering with other institutions to build multi-institution infrastructure

DATA SERVICES NEEDS

- Corral is unique in combining large-scale reliable storage with a complex of data services appropriate to research needs, alongside TACC resources in general
  - Web access – open data publication and publication with minimal protection
  - Access management – both group sharing and HIPAA/FERPA level data protection
  - VM access – many, many custom web and other applications leveraging Corral via NFS
  - Database services – large universe of structured and semi-structured data stores with specialized storage requirements
- Also, Frontera, Stampede, Lonestar, Longhorn, Rustler, etc

NATIONAL CYBERINFRASTRUCTURE AND DATA

- NSF’s XSEDE CI does not provide direct support for most ”data services” needs
  - Many complex reasons for this, but fundamentally XSEDE remains compute-oriented
- Corral provides a critical component for many national-scale data-CI services:
  - CyVerse, nee iPlant (petabyte-scale)
  - Galaxy Bioinformatics Platform (petabyte scale)
  - DesignSafe, DARPA SD2E, other integrated web/compute/data CI projects
  - Arctos and around a dozen other Museum/Archive/Library digitization efforts
  - Many, many DNA/RNA sequence and fMRI datasets with varying user communities
- Storage needs like Open Access page charges in terms of budget (cost and funding mechanism)

BUT DIDN'T YOU SAY IT WAS ABOUT EXPERTISE?

- Collaboration efforts are key especially in national CI efforts
- Many projects involve several components that must interact, including TACC and non-TACC network resources
  - CyVerse includes pieces running in multiple locations
  - Texas Digital Library Chronopolis collaboration includes commercial, TACC, UCSD/SDSC and other resources
- TACC role helping develop deployment plans, decide on resources, deploy cyberinfrastructure, as appropriate

COLLECTIONS COLLABORATIONS - MULTILAYERED

- Simple data hosting – provide persistent web URLs for file objects referenced by external catalog/web front end (TORCH Digitization)
- “Cloud Provider”– both persistent data hosting on Corral and web applications running on TACC VMs, managed by project staff (Symbiota instances)
- Integrated Infrastructure – databases running on specialized hardware, more extensive TACC CI support (Arctos)
- Development Assistance – Full integration of TACC staff into project teams, TACC management of CI resources and deployment (UT Plant Resources Center, Fishes of Texas, IsoBank)

CHANGING HATS - ISOBANK

- Developing a Stable Isotope data repository with partners at Wisconsin, New Mexico, Utah, and New Brunswick
- Rich metadata to include, and link to, museum-provided data sources
  - But our universe of sample sources is much broader than museum specimens
  - Lots of possibility for cross-collection linkage
- Possible to imagine a future world in which multiple "views" of collections specimens are available from the perspective of media type and/or specimen search, and "rabbit holes" lead from one repository to another

HOW TO LEARN MORE OR GET STARTED WITH TACC

- https://www.tacc.utexas.edu – General Information
- https://portal.tacc.utexas.edu – Create TACC accounts, manage resources
- Email: data@tacc.utexas.edu for general data-related queries
- ctjordan@tacc.utexas.edu