Corral: A Texas-scale repository for digital research data

Chris Jordan
Data Management and Collections Group
Texas Advanced Computing Center

UT Research Cyberinfrastructure

- $23 Million allocation from UT Board of Regents
  - Majority of funding to create 10Gb research network
  - $4 Million assigned to deploying a replicated research data repository
- Storage Committee formed to develop requirements and issue RFP to vendors
- RFP completed mid-2011
- Deployment began January 2012
- Research network deployment ongoing

Goals of the Repository

- Provide a highly scalable online resource for both active research data and archived data
- Provide a high-reliability, high-integrity resource to researchers on all UT campuses
- Particular emphasis on support for health institutions and biology-related data
- Support the full range of research scales from desktop to HPC

Technical Architecture

- Two Installations, each with:
  - 4.8PB SATA disk & 300TB SAS disk (DDN)
  - 12 Dell data and service nodes
  - IBM GPFS file system software
- 20GB/sec storage I/O
- Up to 80Gb/sec network I/O
- Data synchronously replicated between sites

Access Mechanisms

- iRODS Data Grid software provides
  - Metadata creation, management and search
  - WebDAV and other interfaces
  - Management mechanism for Web-accessible data
- Simple command-line access (SSH/SFTP)
- Database services (MySQL/Postgres/Oracle)
- Custom web applications (Arctos, XNAT, etc)

Supporting Data Management

- Training on writing and executing data management plans given regularly
- Partner with researchers on grant proposals
- Improving accounting system to support collection-level metadata (policies, etc)
- Automating collection-level policy execution
- Support group, public data sharing

Concerns for Health Data

- Confidentiality, security, privacy
- Importance of large reference collections
- Accelerated evolution of data and metadata standards
- Importance of diverse, often incompatible data types and sources
- Relative novelty of large digital data as a central component of the research

Sustainability/Business Plan

- Provide up to 5TB or storage free initially
- Larger/longer allocations : $250/TB/year
- Some storage set aside for strategic projects
- Likely to request further funding from regents, but goal is to become self-sustaining
- Allocation costs will fund expansion
- Current hardware will run for >3 years, typical media replacement cycle 3-4 years

Long-Term Preservation Issues

- Not an initial goal of the system but highly requested and essential to the institution
- Cost of hardware over generations of media
- Policy tracking and enforcement
- Technical resources for format conversion and/or emulation available at TACC
- Requires researcher investment in metadata
- Mostly an issue of institutional commitment