Intro to Data Management

Chris Jordan
Data Management and Collections Group
Texas Advanced Computing Center
Why Data Management?

- Digital research, above all, creates files
  - Lots of files
- Without a plan, protecting, sharing, and even locating data can be a challenge
- With a plan, researchers can focus on their areas of expertise
  - Many management policies can be automated
  - Replication, open access, cross-collection links
TACC and Data Management

- Data Management has become integral to the conduct of research
  - “Smart grids”, genomics, medical imaging
  - Astronomical imaging, social science, digitization
- Bringing infrastructure and expertise together
- Capacity is never a concern
- Performance (almost) never a concern
- Allows focus on policy and practices
Basic Principles of Data Management

• Think in terms of the whole collection
  – Understand the life cycle of the data beforehand

• Plan for both use and re-use
  – Open Access is always the best choice
  – Access can be subject to embargoes

• Don’t try to do it all yourself!
  – Multi-collection repositories are often available
  – Can make providing access much easier
Life-cycles for data

• Components of the life cycle:
  – Generation for specific purposes
  – Creation of metadata
  – Direct use in research/experimentation
  – Provision of open access
  – Retirement of inaccurate/outmoded data
  – Archival of not immediately useful data
  – Long-term preservation
  – Incorporation into larger repositories
Data Structure

• Think about the entire collection of data to be generated or used in research workflow
• Understand relationships between data of varying types and structures
• Impose structure on data in a consistent way
  – Example: Data in dated directories based on generation time, or in topical directories based on type?
Data Structure 2

• Heterogeneous collections
  – Comprised of both structured data (catalogs) and unstructured data (images/movies/3D)
  – Ideally, catalogs and digitization products should be linked
  – Ideally, these linkages are available via open network mechanisms such as the web
  – Much easier to do this at the time of digitization
“Hidden Knowledge”/Idiosyncracy

• Everyone has their own natural way of organizing data
• Often use “internal conventions” or hidden knowledge in file/directory names
  – Example: Use of dates – generation date? Transfer date? Post-processing date?
• Not specific to digital realm – lab notebooks
• Consistency, transparency, clarity
DM Planning and Execution

- Designate one or more individuals with primary responsibility for data management
- Where possible, partner with experts:
  - Information Scientists, existing repository managers, TACC and similar organizations
- Develop a plan before data is generated
  - Workflows should include data destinations, linkage to databases, etc
Types of Data and Analysis

- Structured vs Unstructured
  - Visualization vs Analysis/Reduction
  - SQL vs NoSQL
- Data format may be related to analysis
- Data may undergo multiple transformations
  - Example: CSV, XML, 4D grids vs. Visualization
- Data may be reduced/subsetted
Structured vs Unstructured Storage

• The most common storage metaphor is the file system, based on hierarchical collections of binary objects without consistent structure - Images, Sound, office documents

• Structured data has different storage needs

• File Systems are tuned for general use, capacity, sequential read/write

• Structured stores are tuned for strided or random I/O, IOPS
“Structured Data”

• Basic Idea: Your data consists of a number of records, themselves composed of elements in a common structure
• You want to be able to extract/compare records based partially on the structure
• Increasing prevalence of “semi-structured” data (e.g. HTML/XML)
• Structured Query Language - SQL/RDBMS
Relational Database Systems

- Metadata, search, filtering and sorting
- Transactional scenarios such as websites
- Ensuring data quality, allowing audit trail
- Hosted database solutions over desktop
  - Automated and routine backup
  - Multiuser and distributed collaboration
  - Single "system of record," no mysterious out of sync copies. Stable, sustainable.
- Mostly but not exclusively SQL-based
Metadata

• Metadata = “data about data”

• Many types of metadata:
  – Technical (size, type)
  – Legal (license, ownership)
  – Format-specific (structure, bit depth, dimensions)
  – Discipline-specific (species, collection time, observation time)

• Metadata facilitates organization, search and re-use
Provenance and Metadata

• Provenance = an “audit trail” for research
• Reproducibility is key to scientific progress
• Without detailed provenance of digital data, how can results be verified?
• Observational data: location and details of circumstances of collection
• Simulation data: detailed information on software and hardware
Metadata Practices

- Metadata (esp Provenance) often cannot be reconstructed after the fact
- Crucial to record metadata as early as possible in the process
- Automate extraction/creation where possible
- Enforce good recording practices for metadata/provenance details
- Software can help with policy enforcement
Archiving Research Data

- Repositories (centralized model)
  - Domain specific and institutional
  - Static data
  - Collection size and functionalities limitations

- Researcher gives access (decentralized model)
  - Issues with maintenance and permanence

- Evolving data storage (semi-centralized)
  - Flexibility to configure/curate your collection
  - Seamless transition between research stages
  - Long-term agreements
What Not to Do

• Do not “shoot first, ask questions later”
• Do not keep only one copy
• Do not go to Fry’s/NewEgg/Best Buy
• Do not use a commercial “cloud” provider as a primary data store
  – Fine for archival copies, costly for access
• Do not use Excel or Access to build a catalog
  – Good for development, bad for stewardship
TACC Resources for Data

- **Corral** – 4PB replicated resource for data collections/management/sharing
- **Stockyard** – 20PB file system for large-scale storage of data used in computation/analysis
- **Ranch** – 160PB Tape archive for long-term storage of relatively low-value data
- **iRODS** – Software “data grid” allowing single namespace spanning many resources
Collections at TACC

• Corral supports both structured and unstructured data stores
• VM Capabilities allow for hosting of websites, applications development, etc
• Collection/Discipline-Specific websites
  – http://www.fishesoftexas.org
  – http://arctos.database.museum
iRODS in Data Management

- Integrated support for metadata and policies
- Numerous APIs: script, compiled, REST
- Web, Command-line, GUI interfaces
- Automated enforcement of rules
  - Metadata extraction
  - Format transformation
- Appropriate for large, complex data collections
Supercomputing for Collections?

- Most collections-related problems are "embarrassingly parallel"
- Image conversion, resizing, OCR, etc scale linearly with added cores/nodes
- Can process tens of thousands of files within hours or days rather than weeks
- Potential for interesting new analysis applications using aggregated collection data
Example: Bioinformatics

- Diverse data types – Structured and Unstructured
- Data sharing critical – openly and within collaboration
- Complex decisions regarding provenance and data retention
  - Sequencer output and derived data products
  - Variable value, huge data sizes
- Well-developed metadata and other standards
  - Perhaps too many of them?
- Planning is key to successful research
  - Infrastructure can be a special challenge
- Many open repositories and group efforts
  - Genbank, TAIR, etc – uncertain sustainability models
- Have your own resources, share data when possible
Example: Arctos/Collections Management

- Arctos hosted entirely at TACC
- Web application with catalog/media linkage/open access capabilities
- Semantic web, export to GBIF, etc
- Many collections in Arctos, including:
  - Museum of Vertebrate Zoology, UC Berkeley
  - Museum of Southwestern Biology, U New Mexico
  - University of Alaska Museum and Herbarium
Example: Simulation Data

- Consider how data furthers research goal
- Software often evolving, community-based, or self-developed
- Output validation could be important
- Plan for validation and analysis
- Consider ease of regenerating output
- Consider time and cost of retention
  - (file-management is effort-intensive)
- “Keep everything” is not a plan
We’re not alone …

- TACC is not the only possible partner
- Similar advanced computing centers exist at many Universities
- Many large-scale repositories that may have relevant expertise and infrastructure
- Aggregator sites facilitate search/use of data
- Data is more valuable the more of it there is, and the easier it is to access
Contacts and Questions

• Chris Jordan – ctjordan@tacc.utexas.edu
• Questions about Data @ TACC?
  – E-mail data@tacc.utexas.edu

• http://www.tacc.utexas.edu
• https://portal.tacc.utexas.edu