Data Management for Digitized Collections

Chris Jordan

Data Management and Collections Group

Texas Advanced Computing Center

Texas Advanced Computing Center

- Serves researchers at the national scale and at UT institutions
- Enabling research through the application of advanced computing technology
- Historically this means supercomputing
- Now, it means large-scale data infrastructure
- Staff expertise as important as “big iron”

Big Iron Highlights

- Stampede – top 10 Supercomputer
  - 10 Petaflops, >100,000 cores, >200TB memory
- Maverick – interactive data analysis and viz
- Corral – 4 petabyte, replicated storage dedicated to research data management
  - >2,000 drives, >20GB I/O per second
- Ranch – 160 petabyte tape archive
  - 20,000 tapes, 2 robotic tape siloes

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

Why Data Management?

- Digitization, above all, creates files
  - Lots of files
- Without a plan, protecting, sharing, and even locating data can be a challenge
- With a plan, collections staff can focus on their areas of expertise
  - Many management policies can be automated
  - Replication, open access, links to specimen records

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 Management for Collections

- Collections have an interesting property:
  - 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

Planning and Execution

- Designate one or more individuals w/ primary responsibility for data management
- Where possible, partner w/ experts:
  - Information Scientists, existing repository managers, TACC and similar organizations
- Develop a plan before digitization begins
  - Digitization workflow should include data destinations, linkage to databases, etc

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 and collections digitization

- Corral supports both structured and unstructured data stores
- VM Capabilities allow for hosting of websites, applications development, etc
- Specify or collection-specific databases
  - http://www.fishesoftexas.org
  - http://www.odonatacentral.org
  - http://www.paleocentral.org

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

TACC and Arctos

- 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

TACC and Specify

- Specify can use external MySQL database
- MySQL can be hosted at TACC
- Images can be hosted at TACC
- Web attachment mechanisms in Specify can link databases to images
- Web-based mechanisms provide location-independence, increased robustness, and transparent scaling

iPlant Collaborative

- Originally funded by NSF to provide Cyberinfrastructure for plant science
- Developed/hosted at TACC and U Arizona
- Now expanding to support additional life science communities, including collections
- Support for large-scale data storage, processing applications, genomics, etc

We’re not alone …

- TACC is not the only possible partner
- Similar advanced computing centers exist at many Universities
- Projects such as iPlant and iDigBio may have relevant expertise and infrastructure
- Data is more valuable the more of it there is, and the easier it is to access
- Cross-collection partnerships are key

Contacts and references

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

- http://arctos.database.museum
- http://www.iplantcollaborative.org
- http://www.tacc.utexas.edu
- https://portal.tacc.utexas.edu