Evolving Collections

Chris Jordan
Data Management and Collections Group
Texas Advanced Computing Center

TACC Mission & Strategy

The mission of the Texas Advanced Computing Center is to enable discoveries that advance science and society through the application of advanced computing technologies.

- To accomplish this mission, TACC provides:
  - Resources and services supporting researchers, educators, developers (science & society applications)
  - Research and development to
    - enhance these technologies and services (software, hardware)
    - apply these technologies in applications
  - Education and outreach
    - Promote awareness of and advocacy for supercomputing
    - Increase participation in advanced computing careers

TACC

A Brief History of TACC

- Established by UT System as an advanced computing center in 1986
- Became part of UT Austin in early 90's
  - Focus on providing mid-level resources in high performance computing
- 'TACC' founded in June, 2001 with ~15 staff
- Today, 80+ staff and growing
- Deployed "***Ranger***" in February 2008
- DMC Group founded, Corral deployed in 2009

TACC

What is the “Data Deluge”?

- Field notes and recordings
- Specimen catalogs and images
- Telescope and Satellite recordings
- Gene and Protein Sequences
- Maps and other geographic data
- Specialized sensor outputs
- … And nobody has a good place to put it

A Collections-Driven Approach

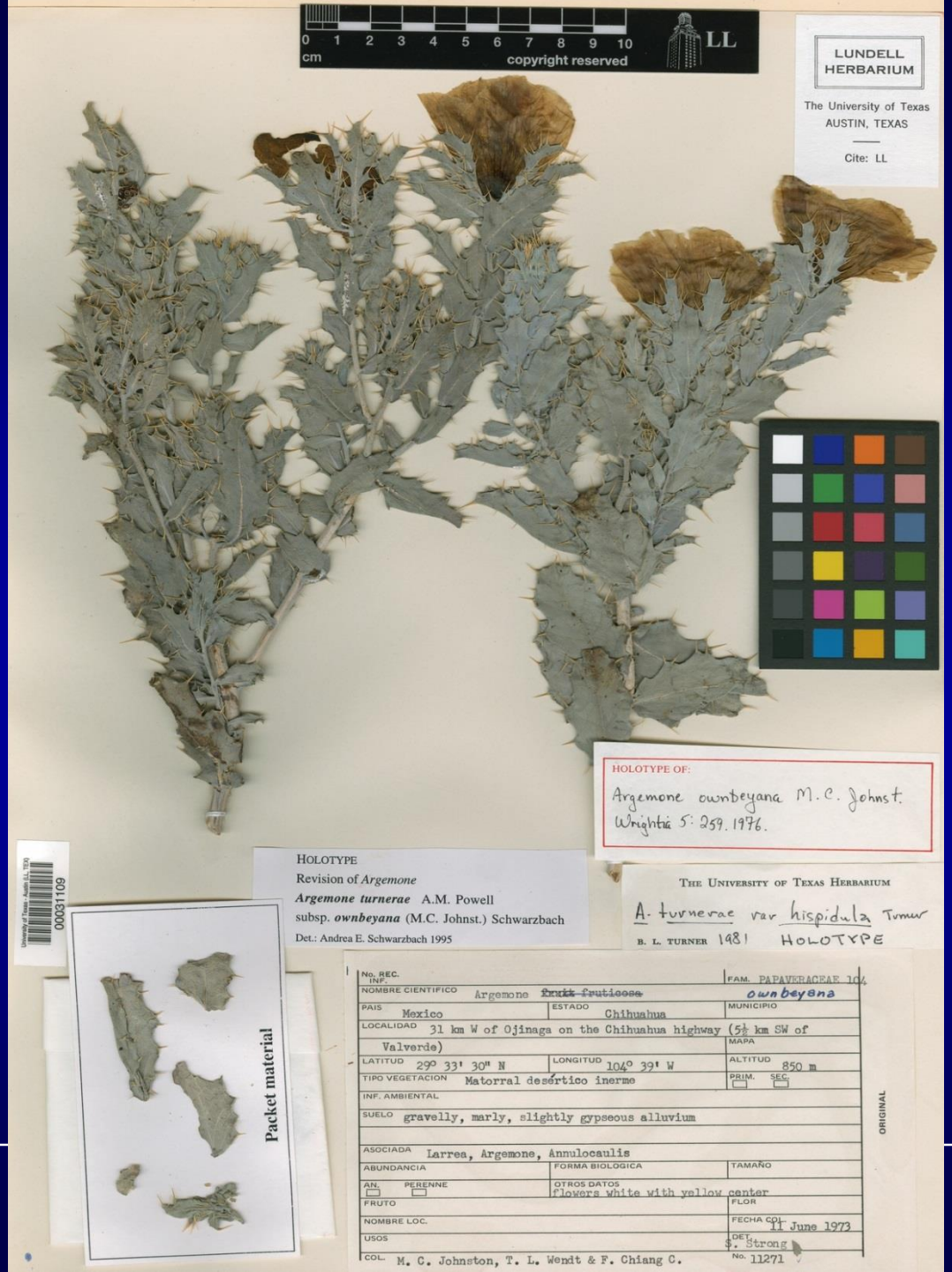
- Research practices around data are still developing (and will continue to)
- Broad array of data needs and types
- Little practical experience or documentation
- Solutions are driven by data and practice
- Evolving over time through collaborations

Goals in Collaboration

- Minimize intrusiveness of technology
- Consolidate data onto reliable systems
- Incorporate robust data management into existing/improved workflows
- Enable improved control and discoverability
- Enable data sharing and public access

Plant Resources Center

- >1 Million Pressed Plant specimens
- 4 separate “collections”, 25% digitized
- 10s of thousands of high-quality images
  - 5% Digitally Imaged
- “Distributed” storage on desktops
- Inconsistent schemas & taxonomies


0 1 2 3 4 5 6 7 8 9 10
cm
copyright reserved
LL
LUNDELL HERBARIUM
The University of Texas
AUSTIN, TEXAS
Cite: LL
HOLOTYPE OF:
Argemone ownbeyana M.C. Johnst.
Wrightia 5: 259. 1976.
HOLOTYPE
Revision of Argemone
Argemone turnerae A.M. Powell
subsp. ownbeyana (M.C. Johnst.) Schwarzbach
Det.: Andrea E. Schwarzbach 1995
00031109
Packet material
THE UNIVERSITY OF TEXAS HERBARIUM
A. turnerae var hispidula Turner
B. L. TURNER 1981 HOLOTYPE
No. REC. INF.
FAM. PAPAVERACEAE 104
NOMBRE CIENTIFICO Argemone ~~pruik fruticosa~~ ownbeyana
PAIS Mexico
ESTADO Chihuahua
MUNICIPIO
LOCALIDAD 31 km W of Ojinaga on the Chihuahua highway (5½ km SW of Valverde)
MAPA
LATITUD 29° 33' 30" N
LONGITUD 104° 39' W
ALTITUD 850 m
TIPO VEGETACION Matorral desértico inerme
PRIM. SEC.
INF. AMBIENTAL
SUELO gravelly, marly, slightly gypseous alluvium
ASOCIADA Larrea, Argemone, Annulocaulis
ABUNDANCIA
FORMA BIOLOGICA
TAMAÑO
AN. PERENNE
OTROS DATOS flowers white with yellow center
FRUTO
FLOR
NOMBRE LOC.
FECHA COL. 11 June 1973
USOS
DET. S. Strong
COL. M. C. Johnston, T. L. Wendt & F. Chiang C.
No. 11271
ORIGINAL

PRC & TACC

- Moving all images to TACC
- TACC Conversion to JPEG / Thumbnail
- Database Consolidation
  - Tomislav Urban (TACC) and Tom Wendt(PRC)
  - New, consistent schema for all specimens
  - Normalized Taxonomy etc
  - Platform for collections curation
  - DarwinCore for data sharing

Museum of Vertebrate Zoology

- UC Berkeley - Ornithology Collection
  - Eggs, historical photos, field notes, audio recordings, etc
  - >100 Thousand diverse digital objects
  - Arctos multi-collection web/database for cataloging (Alaska Data Center)
  - Need terabytes of web-accessible, highly reliable storage

No. 5 Subject Extreme west end of Salton Sea

Date April 19, 1908 Locality Salton Sea, near Mecca, Cal. Photographer J. Grinnell

~~no neg.~~

~~Copy neg 2/2000~~

University of California
Museum of Vertebrate Zoology

MVZ Workflow

- SSH/SCP “Drop box” for simple upload
- TACC scripts for image processing and ingest
  - Format conversion, resizing, checksumming
  - Automated replication to archive
- Consistent, predictable URLs
- URLs linked into Arctos database

Institute of Classical Archeology

- Excavations in Chersonesos and Southern Italy over last 20 years
- Geographic surveys, field notes, objects, images, and publications
- “Distributed” Storage in unspecified locations
- Want to link information, artifacts, and publications on an ongoing basis

ICA & TACC

- ~1 year of collaborative effort understanding metadata, workflows, etc
  - Maria Esteva, Tomislav Urban, Adam Rabinowitz, Jessica Trelogan
  - “Archive” processes will be simultaneous with “Working” processes
  - Working objects will have full provenance, descriptive metadata, and numerous versions
  - Publications and web pages will be linked to the archive

Broad Infrastructure

- Corral data-intensive applications facility
- Ranch long-term tape archive
- Visualization and Analysis facilities
- Policies for the data life cycle
- Practices for effective data management
- TAS Collections Catalog

Technical Infrastructure

- Corral – Petabyte disk resource
  - 16 Data and Application Servers
  - Interactive, web, database, & other services
- Ranch – Tape archive system
  - 10 Petabyte capacity, expanding to 60PB
- Offsite replication - Indiana University, SDSC
- iRODS Data Management
  - Links all storage systems, provides policy tools

Human and Policy Infrastructure

- Group developing expertise in discipline-specific practices, standards & formats
- Patterns for data management
- Policies for
  - Reliability
  - Availability
  - Ownership, Retention and Succession
- Goal is to automate policy enforcement

Collections Catalog

- Extension of TACC Accounting System (TAS)
- Schemas for description, control, and contact information
- Tracking of policies and procedures
- Will link to iRODS catalog for collections tracking and policy enforcement purposes
- Provides the basis for automation of access, management, & preservation actions

Future Plans

- Corral Expansion
  - Up to 2.4 Petabytes
  - Expanded server capacity
  - High Availability provisioning
- Replication with TeraGrid/other resources
- Formal policy “menu” linked to TAS
- More collections, more disciplines
- “Spinning Archives” low-cost reliable disk

Thanks

- MVZ – http://*mvz.berkeley.edu/*
- PRC - http://www.biosci.utexas.edu/prc/
- ICA - http://www.utexas.edu/research/ica/
- TACC – http://www.tacc.utexas.edu

- Data Management and Collections Group
  - Maria Esteva, Tomislav Urban, David Walling