# Data Migration from Excel: A Brief Introduction to Arctos

Dusty McDonald, UAM

Carla Cicero, MVZ

iDigBio Workshop

Vertebrate Digitization II

Berkeley, California

4-6 April 2016

# What Is Arctos?

Collection Information Management System for managing data on:

- Specimens
  - Basic “label data”
  - Usage: Projects → Loans → Publications → Genbank
  - Interactions: host/parasite, predator/prey, parent/offspring…
- Agents
  - People
  - Agencies
  - Groups
- Places and Events
  - Descriptive
  - Spatial
  - Temporal
- Taxonomies
- Media
- Other Stuff

# Arctos is…

- Powerful Oracle Relational Database
- Comprehensive set of web applications
- Interaction with external services
- Community

# Two Key Features

- Normalization
  - Eliminate redundancy
- Standardization
  - Don’t say the same thing multiple ways

# Label Data Are Core!

For every standardized term, there is an accompanying verbatim, e.g.

| | | |
|---|---|---|
| Taxonomy | ← → | Identification |
| Geography | ← → | Verbatim Locality |
| Events | ← → | Verbatim Date |
| Agents | ← → | Verbatim Collectors |

Verbatim data have structure too

- Remarks field for every object
- Datatype Attributes

# Datatype

Free text

- “ad.”
- “skull ossified”
- “sngl-layrd sk.”

Categorical

- young adult
- adult

Numeric

- 12 years
- 3 days

# Goals of Digitization

Digitization should allow the data to DO STUFF

- No pre-defined boundaries
- Interact with other data
- Be discoverable
- Add value to physical specimens

# Tools

No person, no matter how meticulous, can consistently enter data at any useful scale

Tools to enforce consistency are critical

- Pick, don't spell
- Enforce Rules on Datatype
  - Latitude can't be > 90°
  - "Dates" which are not (Feb 31)
  - Non-numeric data with units
  - Arbitrary arrangement of terms (geography, higher taxonomy)

# The Excel Challenge

How do you migrate non-standardized verbatim values to standardized, structured data for easy discovery?

Let’s look at Agents (People) as an example of how this works in Arctos…

# The View from Excel

- C. L. Parker
- Carolyn Parker, Page Specner, & Stacy Studebaker
- Carolyn Parker, Page Spencer, & Stacy Studebaker
- Carolyn Parker, Page Spencer, Stacy Studebaker
- Carolyn Parker, Page Spenver, & Stacy Studebaker
- Carophy Parker, Page Spencer, & Stacy Studebaker
- Carophyn Parker, Page Spencer & Stacy Studebaker
- Stacy Studebaker & Carolyn Parker
- Stacy Studebaker and Carolyn Parker
- Stacy Studebaker & Caroyln Parker
- Stacy Studebaker & Caryolyn Parker
- Stacy Studebaker, Carolyn Parker
- `
- 53

# The View from Arctos: One Agent

Agent Names:

- Carolyn L. Parker (preferred)
- Carolyn Parker (aka)
- carolyn (login)
- C. L. Parker (initials plus last)
- Parker (last name)
- L. (middle name)
- Parker, C. L. (last plus initials)
- Carolyn (first name)
- CLP (initials)
- Parker, Carolyn L. (aka)

# Agent Relationships

Carolyn L. Parker is

- Not the same as Carolyn R. Parker
- Harold Parker is sibling of
- C. L. Parker is bad duplicate of

# Agent Activity

- Collected or Prepared specimens:
  - 1 UAM:EH specimens
  - 7 UAM:Alg specimens
  - 1406 KWP:Ento specimens
  - 1 UAM:Bird specimens
  - 1 UAM:Ento specimens
  - 17001 UAM:Herb specimens
  - 9 UAM:Mamm specimens
  - 255 UAMb:Herb specimens
- Projects
  - U.S. Forest Service-Alaska Region
  - Yupik Ethnobotany Project
  - Three Parameters Plus,Inc.-2012 Annex Creek Botanical Survey
  - Arctic Alaska Network Inventory and Monitoring vascular plant survey.
  - Publications Ihsan Ali Al-Shehbaz, Jason R. Grant, Robert Lipkin, David F. Murray, Carolyn Parker. 2007. Parrya nauruaq (Brassicaceae), a new species from Alaska. Novon 17:275-278.
    - 9 citations
  - Matthew J. Wooller, Grant D. Zazula, Mary Edwards, Duane G. Froese, Carolyn Parker, Bruce Bennett. . Stable carbon isotope compositions of eastern Beringian grasses and sedges: Investigating their potential as paleoenvironmental indicators. Acta Chiropterologica 39(2):318-331.
    - 0 citations
  - Matthew J. Wooller, Grant D. Zazula, Mary Edwards, Duane G. Froese, Carolyn Parker, Bruce Bennett. 2007. Stable carbon isotope compositions of eastern Beringian grasses and sedges: Investigating their potential as paleoenvironmental indicators. Arctic, Antarctic, and Alpine Research 39(2):318-331.
    - 140 citations

# How Do We Get There?
# The Arctos Bulkloader

- All Arctos data entry goes through the Bulkloader
  - Data Entry screen, small imports by Curators, new collections, everything!
- Giant flattened table, accepts CSV
- Arctos provides a builder to create templates
- Can write to other bulkloaders
- New collection imports tested to ~400K specimens, no apparent limitations

# Data Entry

- Writes to Arctos Bulkloader
- Can pull data (eg, “seed” parasite record with host information)
- Can write to related tables (extra collectors, attributes, etc.)
- Can selectively carry data from previous record
- Fully customizable by collection type, collection, user
- API-based

# Pre-Bulkloader

- Cleans data based on structure of bulkloader
- Extracts unique values of controlled data
- Creates lookup files
- Repatriates cleaned data
- Pushes to bulkloader

# Pre-Bulkloader

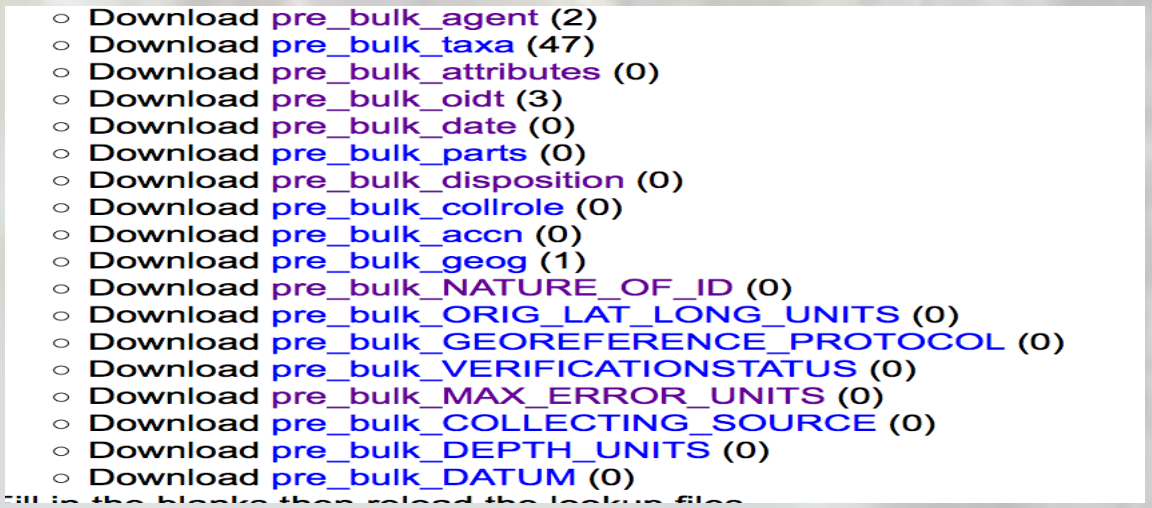
Relies on a series of Pulls from existing data, e.g.

- Agents
  - Name variations (Bob ← → Robert)
  - Non-preferred (maiden/married, AKA, native character sets [*e.g.*, Cyrillic])

- Geography
  - Standardized descriptive data (“Russia”)
  - Less-standardized “search terms” (“Росси́я”, “Russian Federation,” “Росси́йская Федера́ция”)
  - Wikipedia

# Pre-Bulkloader: Demo

- Data for new collection (~3K eggs)
- Assembled into Bulkloader format by collection
- Upload to Pre-Bulkloader, set to check, get list of unrecognized controlled data



Download a table, fill in the blanks:

| | A | B |
|---|---|---|
| | AGENT_NAME | SHOULDBE |
| | Adam Leavitt | |
| | Charles D Brower | |
| | | |

# Check Arctos for existing matches

Collecting Summary

| Collection | Earliest Date | Latest Date | NumberSpecimens |
|---|---|---|---|
| UAM:Bird | 1963-08-06 | 1963-08-06 | 1 |

Rank

save all changes

Edit Agent

**Preferred Name**

A. Leavitt

**Agent Type**

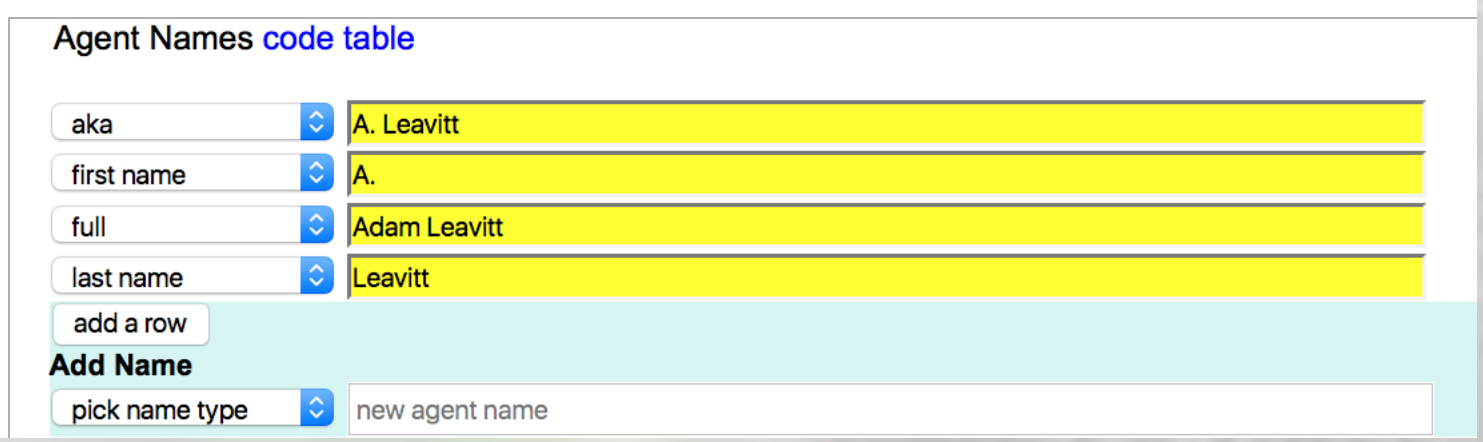
person Define

**Agent Remark**

Bird collection agent.

Looks promising, add a name and remove the “missing” agent from the pre-bulkload lookup

New record will load, existing data improved by additional context



# Mis-match (due to punctuation) corrected in the lookup

| | A | B |
|---|---|---|
| 1 | AGENT_NAME | SHOULDBE |
| 2 | Charles D Brower | Charles D. Brower |
| 3 | | |

# Re-upload lookup files


Results are pushed to all flattened fields

- Collectors
- Attribute determiners
- Event determiners
- Etc.

Cleaned data are re-checked, exported to standard bulkloader, loaded

Currrent state of pre-bulkloader:

| enteredby | numrecs | loaded |
|---|---|---|
| ekrimmel | 3116 | final_check_pass |

What used to be a very technical and error-prone exercise often taking weeks is now a matter of minutes or hours

# Summary

- Have clear goals, organize data accordingly
- Choose appropriate tools
- Authority data is critical
- Understand limitations

Arctos
Collaborative Collection Management Solution

Questions?
dustymc@gmail.com