Digitization of Paleontology Collections: An Overview

Chris Norris
Division of Vertebrate Paleontology
Yale Peabody Museum
What Do We Mean By Digitization?

“Collections digitization is defined broadly to include transcription into electronic format of various types of data associated with specimens, the capture of digital images of specimens, and the georeferencing of specimen-collection localities.”

A Strategic Plan for Establishing a Network Integrated Biocollections Alliance, 2010
Why Digitize?

A significant fraction of the “value” of a natural history specimen is represented by its associated data.

Digitization mobilizes these data, making them available for a wide range of uses.

Mass mobilization offers the potential for far-ranging analyses of collections data.
“Big Data”

Things that can be done at a large scale that cannot be done at a smaller one

Enables you to extract new insights or create new forms of value

Defining features:

- **N=all**: analyze all data rather than a subset
- **Messy**: more data means less precision needed
- **What not Why**: stresses correlation, not causality

Techniques could be applied to collections data if more of them were digitized

Aim

To create… “an inclusive, vibrant, partnership of US biological collections that collectively will document the nation’s biodiversity resources and create a dynamic electronic resource that will serve the country’s needs in answering critical questions about the environment, human health, biosecurity, commerce, and the biological sciences.”

A Strategic Plan for Establishing a Network Integrated Biocollections Alliance, 2010
Ideally...

- “Object-Image-Data”
  - Remove specimen from collection
  - Image specimen
  - Extract label data using OCR
  - Upload image and data to web
  - Repeat many times

Source: http://www.huh.harvard.edu/collections/gray.html
Paleontology presents some additional challenges...

“Collections digitization is defined broadly to include transcription into electronic format of various types of data associated with specimens, the capture of digital images of specimens, and the georeferencing of specimen-collection localities.”
Data Transcription

- Various sources
  - Specimen label
  - Catalog card
  - Field notes
  - Written on specimen

- Variable quality
  - Typed
  - Handwritten
  - More or less complete

- Workflow has to accommodate this

- O-I-D may only be a start
Imaging

- Emphasis on imaging comes in part from O-I-D

  - If O-I-D is less effective for paleo specimens, should you still image?

- Yes

  - Documentation
  - Validation
  - Public access/education
  - Research
But there are challenges

- Variability in size: microfossils – dinosaurs
- Variability in preservation
- Different diagnostic characters
- Difficulty lighting
- Specimens may contain many parts
Georeferencing

- Has a temporal as well as spatial component
- Stratigraphic data may be incomplete or lacking
- Multiple data sources
- Very time-intensive
- Limitations on release of site-specific data
- But… is critical to data utility
Workflow design is critical

- How will you integrate different sources of data?
- How will specimens be staged to minimize issues of size, preservation, etc?
- How can georeferencing be organized to bring the maximum number of specimens on-line as quickly as possible?
- How can digitization be coordinated with other ongoing collections activities?
Some things to consider…

- Imaging specimens or imaging drawers?

- Imaging types or imaging unpublished specimens?

- Imaging in great detail (e.g. CT scanning) or imaging in great numbers?

- Digitizing ledgers vs. digitizing specimens?

- Investing in technology vs. investing in people?