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

Mayer-Schönberger, V. & Cukier, K. 2013. *Big Data: A Revolution That Will Transform How We Live, Work & Think*. New York, Houghton Miflin Harcourt

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…

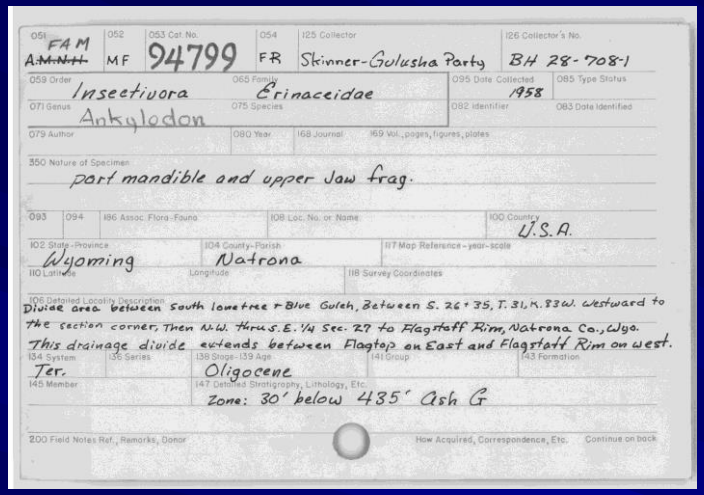
- "Object-Image-Data"
  - Remove specimen from collection
  - Image specimen
  - Extract label data using OCR
  - Upload image and data to web
  - Repeat *many* times

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



051 ~~A.M.N.H.~~ FAM | 052 MF | 053 Cat. No. 94799 | 054 FR | 125 Collector Skinner-Gulusha Party | 126 Collector's No. BH 28-708-1

059 Order Insectivora | 065 Family Erinaceidae | 095 Date Collected 1958 | 085 Type Status

071 Genus Ankylodon | 075 Species | 082 Identifier | 083 Date Identified

079 Author | 080 Year | 168 Journal | 169 Vol., pages, figures, plates

350 Nature of Specimen: part mandible and upper jaw frag.

093 | 094 | 186 Assoc. Flora-Fauna | 108 Loc. No. or Name | 100 Country U.S.A.

102 State-Province Wyoming | 104 County-Parish Natrona | 117 Map Reference - year - scale

110 Latitude | Longitude | 118 Survey Coordinates

120 Detailed Locality Description: Divide area between South Ionetree & Blue Gulch, Between S. 26 & 35, T. 31, R. 83W. Westward to the section corner, Then N.W. thru S.E. 1/4 Sec. 27 to Flagstaff Rim, Natrona Co., Wyo. This drainage divide extends between Flagtop on East and Flagstaff Rim on west.

134 System Ter. | 135 Series | 138 Stage - 139 Age Oligocene | 141 Group | 143 Formation

145 Member | 147 Detailed Stratigraphy, Lithology, Etc. Zone: 30' below 435' Ash G

200 Field Notes Ref., Remarks, Donor | How Acquired, Correspondence, Etc. | Continue on back

- 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

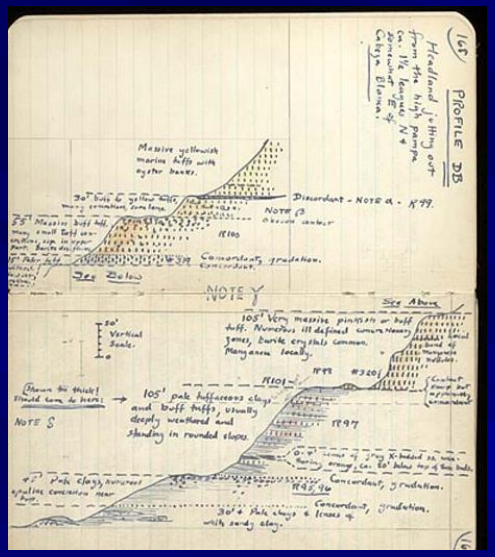
Imaging

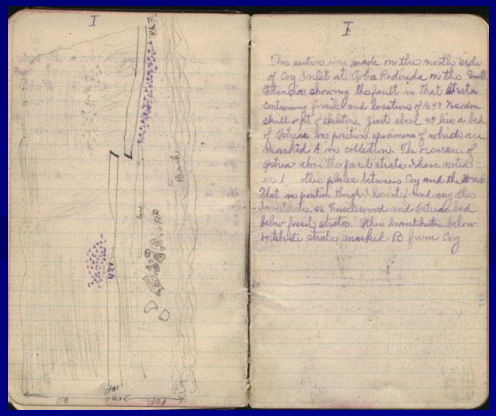
- Emphasis on imaging comes in part from O-I-D
  - If O-I-D is less effective for paleo specimens, should you still image?
- Yes
  - Documentation
  - Validation
  - Public access/education
  - Research

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?
  - See Balke et al, 2013. *Frontiers in Zoology* 2013, **10**:55 doi:10.1186/1742-9994-10-55
- 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?