DROID2: Digitizing Pinned Insect Collections

>>> Getting Started <<<

Margaret Thayer & Petra Sierwald

Field Museum of Natural History

Scope

Pinned insect collections
contain the largest number of individually-labeled specimens of any biological specimen collection type –
probably **billions** worldwide

Why and how do we digitize them?

Why Digitize?

- Mobilize (label) data for better access
  - Research use (e.g., monography, mapping distributional changes)
  - Facilitate public understanding (e.g., area or regional checklists)
- Images
  - Document diversity – for researchers & public
  - Document type condition
  - Facilitate remote access for identification (2-way)
- Facilitate collection management

Kinds of Data Captured

- Species-level inventories
  - Collection management (tray labeling) / curation
  - Flag presence of types
  - Share summary data via web (not shareable pre-EMu)
- Types
  - Traditional: label, curate, safeguard, publish lists
  - New: digitize data ($\rightarrow$ dynamic lists), type images, share detailed data via web
- Specimen-level data capture (non-types)
  - Collection management: loan tracking, more detailed records of holdings
  - Facilitate local and remote research (directly or via aggregating portals)

Methods of Data Capture

- Keyboard entry from labels
- Import from existing digital sources
- Voice recognition software supplementing kb
- Keyboard or voice entry from label images

Specimen-level Data Capture
Start-up costs (1)

- Software
  - Proprietary (e.g., EMu) licenses and custom development costly
  - Free software (e.g., Specify) not fully cost-free: local setup/customization, staff training, workflow development
- Clean existing (legacy) data for import when changing systems
- Institutional IT staff
  - Software and database setup & support, including web serving of data and images, backup

Specimen-level Data Capture
Start-up costs (2)

- Hardware
  - Extra computers
  - Imaging equipment if not available
- Grant writing to hire data entry/imaging staff
- Pre-data-entry curation (staff time)

Specimen-level Data Capture
Continuing institutional support (1)

- Proprietary software: licenses, expenses for further development
- Free software: some continuing costs (implementing upgrades – just staff time?)
- Ongoing staff training (upgrades, new staff); need computer-fluent people
- IT staff
  - Software and database maintenance
  - Data and image backup
  - Support for web delivery of data and images, sharing w/portals

Specimen-level Data Capture
Continuing institutional support (2)

- Pre-data-entry curation for new data entry (staff time)
- Data entry (grant- or institution-supported, or volunteers)
- Cleaning new data (data-entry and/or other staff)
- Computer and imaging equipment
  - Hardware maintenance, upgrades, replacement
  - Software upgrades
  - Data storage (ever-increasing)

Specimen-level Data Capture
>> Benefits <<

- ## Research
  - Monitoring range changes, tracking invasive species, niche modeling
  - Documenting regional biological hotspots
  - Assigning GUIDs facilitates re-study of individual specimens
- ## Collection Management & Loans
  - Online specimen data & images (also collection inventories) streamline inquiries; remote users can sometimes fulfill their needs
  - Feeding data to portals (e.g., GBIF, Symbiota) raises institutional profile, promotes use of data and collection
  - Get remote help with IDs of images
  - Facilitates some annual reports
  - Simplifies loan processing (esp. with bar codes or other GUIDs)
  - Reduces shipping costs, enhances safety of type specimens if images suffice
  - Prospective data capture and label generation work together
  - Enables new things more than it saves time

Specimen-level Data Capture
>> Obstacles <<

- ## High Initial Costs / Low <u>Immediate</u> Benefits
  - Continuing data entry (and <u>updating records</u>) requires more, not fewer, people
  - Existing/new data require cleaning/proofing; imports take staff time

- ## Research
  - Grossly underdeveloped benefits so far
  - Value/benefit not immediately visible, especially to administration/board/prospective funders

- ## Collection Management, Loans
  - Online serving of data and images requires storage and IT expertise, costs
  - Enables – even demands – doing new things more than it saves time

Outlook (1)

- Need to increase data capture rates greatly
  - Automate data capture
  - Image labels: less handling of specimens
  - Keep labels on pins as much as possible
  - Crowd sourcing of data entry from label images – but need careful planning

Outlook (2)

- Collection-based researchers clearly see need for and benefits of digitization

- Research use of digitized data grossly underdeveloped

- Need to focus more attention on research use & benefits to convince a broader audience?

  - Seek out specific collection-based examples from other taxa?

  - Make digitized collection data and images more useful to non-specialists, increase browsing options, make images and data more available for other uses

Thanks to:

- Initial funding and impetus toward FMNH shared database from Bob Martin, then-FMNH Provost
- IMLS, NSF, and Mellon Foundation funded various parts of FMNH's initial EMu conversion & implementation (2002–) and further developments

Questions or Discussion?