Digitization at ASNHC: A SMALL COLLECTIONS APPROACH
Overview of Small Collections

ASNHC Workflow

Helpful Tools & Resources

digitization at ASNHC: a small collections approach
Overview of Small Collections

digitization at ASNHC: a small collections approach
1923 by A. Brazier Howell

- 37 public and 40 private collections in North America (U.S. & Canada only, sorry Mexico)
- ½ of all specimens at Smithsonian

“The larger museums are relatively sufficient unto themselves; but the smaller ones, as well as unattached individuals, are often handicapped…it is hoped the appearance of this list will have is the stimulation of the interest of the small collector, the beginner, and the amateur.”

<table>
<thead>
<tr>
<th>Year</th>
<th>No. Collections</th>
<th>Growth of Specimens Added Between Surveys</th>
<th>% Growth</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1945</td>
<td>297</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1961</td>
<td>307</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1973</td>
<td>388</td>
<td>1963-1973: 900,000 specimens = 90,000/yr</td>
<td>57%</td>
<td></td>
</tr>
<tr>
<td>1983</td>
<td>474</td>
<td>1973-1983: 600,000 specimens = 60,000/yr</td>
<td>24%</td>
<td>First mention of digitization</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3,136,557 specimens</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1996</td>
<td>329</td>
<td>1983-1995: 670,000 specimens = 56,000/yr</td>
<td>21%</td>
<td>Increase in number of collections with 10,000 or less</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3,806,557 specimens</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4,124,065 specimens</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Size of Mammal Collections in United States

307 Collections
3,886,071 specimens

- 43% <1,000
- 37% 1,001-10,000
- 15% 10,001-50,000
- 2% 50,001-100,000
- 3% >100,000

ASM Systematics Committee, 2007
Location of Collections in United States

ASM Systematics Committee, 2007

307 Collections
3,886,071 specimens

104 databased 34%
41 accessible to public 13%

Map from ASM, 1997
Collections 2015
Comparison of Portals

iDigBio
• 79 collections
• 3,341,640 specimens

ASM Systematics Committee Survey, 2007
• 307 Collections
• 3,886,071 specimens

Over half million specimens left
Mostly from small collections
LIKELY MORE

225+ Collections
There are many large natural history collections, but it’s really the small collections that provide detailed knowledge of the flora and fauna of local and regional communities.

- Hilary Moore, Importance of Small Museum Collections

DIGITIZATION INCREASES VALUE OF YOUR COLLECTION
Advocating your collection = Advocating your job
Concentration of Records vs Location of Collections

iDigBio, 2015

ASM, 1997
Small Collections

Big Benefits

Successfully Curating Smaller Herbaria and Natural History Collections in Academic Settings, Neil Snow, *BioScience* 2005

CHALLENGES

- Limited Budget
- Limited Time
- Limited Staff
- Lack of IT Support
digitization at ASNHC: a small collections approach
Angelo State Natural History Collections

- 5 Collections
  - Mammalogy
  - Ornithology
  - Herpetology
  - Herbarium
  - Frozen Tissues

  ~ 150,000 objects

- 5 Curators – all full-time faculty with no formal time allocated toward collections
- 1 Full-time Collections Manager – hired 2013
Mammalogy Collection

- Began in 1968 (3 curators)
- 4th largest in Texas
- Nearly 50% of collection is from Texas. Also has concentrations from Mexico, Guyana, and Ecuador

- Mostly databased
- Specify
- Approximately 18,000

CURATOR

Dr. Robert C. Dowler
Digitization Workflow for Mammalogy Collection at ASNHC

ENTIRE PROCESS

INTEGRATING NEW SPECIMENS

STARTING WITH DIGITIZED DATA
Helpful Tools & Resources

digitization at ASNHC: a small collections approach
Data Cleaning Tools

- Power tool for working with messy data, cleaning it up, transforming it from one format into another, extending it with web services, and linking it databases like Freebase
- Not a web service, but a desktop application, interact with it on web browser

Data can be in spreadsheet or other format: TSV, CSV, *SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, Wiki markup, and Google Data documents

OpenRefine Core

**Google Refine 2.5 - Stable version**

- **Windows kit**, Download, unzip, and double-click on `google-refine.exe`. If you’re having issues with the above, try double-clicking on `refine.bat` instead.
- **Mac kit**, Download, open, drag icon into the Applications folder and double click on it. **NOTE:** If you have issues installing Refine on Mac, please refer to issue 590 - OpenRefine 2.5 for mac support java 6 and 7 only
- **Linux kit**, Download, extract, then type `./refine` to start. **NOTE:** OpenRefine 2.5 for linux support java 6 and 7 only

http://openrefine.org/download.html
Data Cleaning Tools

Configuring the File

- Special Characters (ä, ü) - choose “UTF-8” encoding
- Custom radio button if your file is tab-delimited, or has a delimiter different than comma
- Discard rows, ignore rows, parse data, in this initial step to create a project.
- NOTICE: “Parse cell text into numbers, dates, …” Untick this box unless you want this program to convert your data into numeric or date format. If you want your dates verbatim, and the significant digits of your numbers unchanged, untick this box.
Data Cleaning Tools

Text Facet shows you total number of variants, total number of each variant, and can be used to identify inconsistencies.
Data Cleaning Tools

Editing capabilities let you identify and clean up typos that exist.
Data Cleaning Tools

Cluster feature will identify fields with similar data and provide option to do editing.

Always maintain a field to capture your locality in original or ‘verbatim’ form before editing.
Data Cleaning Tools

If you make mistakes... no problem!
Georeferencing Tools

Collaborative Georeferencing

Guidelines

- Train the Trainers
- Georeferencing Working Group
Database

Specify 6 ➔ SpCloud ➔ Specify 7
Database
Angelo State Natural History Collections
Mammalogy

Curator Robert C. Dowler
Collection Manager Marcia A. Revelez
Created On Jun 12, 2015

GEOGRAPHY
TAXONOMY
STATISTICS
SUMMARY
Database Tool

GEOGRAPHY
Database Tool

STATISTICS
Inventories

- Utilize Database from smartphone
- Be in the cases, not anchored to computer
- Create Inventories, Loans, Reports, etc.
Digitization Workflow for Mammalogy Collection at ASNHC

**ENTIRE PROCESS**

**INTEGRATING NEW SPECIMENS**
- Create Data Collection Standards
- Train Students & Curators
- Research & Collecting
- Quality Check & Address Problem Data
- Cataloging Specimens
- Enter Data into Specify Workbench

**STARTING WITH DIGITIZED DATA**
- Data located on server
- Create Data Standards
- Export Data from Database
- Data Cleaning via Google Refine
- Quality Check & Address Problem Data
- Georeference
- Merge Georeferenced Data & Cleaned Data
- Specify 6
- Specify 7
- Specify Insight
- Cloud based Usage

New Data Cleaning Tool
Kurator: Towards Data Curation Workflows for Mere Mortals
An extensible, open-source workflow platform for users & makers of data curation tools

B. Ludäscher  J. Hanken  D. Lowery  J.A. Macklin
T. McPhillips  P.J. Morris  R.A. Morris  T. Song

Towards user-definable, semi-automated workflows for curating biodiversity data

P.J. Morris  R.A. Morris  B. Ludäscher
D. Lowery  J.A. Macklin
T. Song  T. McPhillips  J. Hanken

Webinar hosted by iDigBio
Integrated Digitized Biocollections
Kurator: Kepler Data Curation Package

- Software in development – mechanism for quality control of biodiversity data

- Problems with Data and Metadata Quality
  - Lat/Long transposition and other georeferencing issues
  - Scientific Names (spelling errors, out-of-date nomenclature, etc.)
  - Data entry/creation can be full of errors associated with data conversions and transformations, schema mappings, or human error

- Problems Kurator is Attempting to Solve
  - Detect and flag data quality issues
  - Provide opportunity for repair
  - Employ workflow (semi-)automation

GOALS: 1) get data quality control tools out into the hands of people who are vested in the quality of their own data, and 2) get feedback from them to improve the capabilities of our tools
3 Key Questions to Test Quality of Data

- **What?** – scientific name validation
- **When?** – collecting event date validation checks a collector’s life span vs. date collected
- **Where?** – georeference validator

Data cleaning workflow:

1. Load Data
2. Check scientific name
3. Check basisOfRecord
4. Check date collected
5. Check lat/long
6. Geolocate
7. Write out results
Data Cleaning Tools

Taxonomic Validation
Data Cleaning Tools

Kurator

Georeference Validation
<table>
<thead>
<tr>
<th>Occurrence Id</th>
<th>Event Date</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>01-01-2023</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>02-02-2023</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>03-03-2023</td>
<td></td>
</tr>
</tbody>
</table>

**Data Cleaning Tools**

**Kurator**

**Collecting Event Date Validation**
Resources

Websites, Listservs, Meetings, Symposia, Workshops, YouTube videos, documents, powerpoint presentations, recordings, and so much more!
Resources

Students

Get Involved

EDUCATE AT EVERY OPPORTUNITY

Training the next generation of stewards
Benefits of Student Involvement

**Become advocates for natural history collections**

**GIVE**
- Critical to the successful execution of the collections goals
- Carry out the high quality of research conducted by faculty and staff
- Contribute to the university’s mission

**TAKE**
- Gain a better understanding and appreciation for best practices, natural history museums, and the greater scientific community
- Gain an appreciation for the area flora & fauna of the Concho Valley and beyond – understanding in biodiversity, conservation, climate change, etc.
- Obtain skills in research, work ethic, time management, accuracy, responsibility, and professional courtesy
- Inspiration in future career options
But Wait

There’s More...

digitization at ASNHC: a small collections approach
Advance Digitization

- Genbank
- Zoo Record Sheets
- Texas Department of Health Services
- Literature Citations
- Images
- Video
Advance Digitization

- Tissue Data pages (ASK #)
- Journals
- Catalogs
- Specimen (with tag data)
Acknowledgments

National Science Foundation’s Collections in Support of Biological Research - Modernization and Digitization of the Angelo State Natural History Collections” (1203363)

Brian J. May, President, Angelo State University
Russell Wilke, Chair, Dept. of Biology, Angelo State University
Robert Dowler, Loren Ammerman, Bonnie Amos, Terry Maxwell, Ben Skipper, and Mike Dixon, Curators, ASNHC

Deb Paul, Gil Nelson, Kevin Love, Joanna McCaffrey

James Beach, Teresa Miller, Andy Bentley

Paul Morris, Bertram Ludascher, James Macklin, D. Lowery, R. A. Morris, T. Song, T. McPhilips, and J. Hanken

Cody Thompson, American Society of Mammalogists