Undertaking digitization in the NMNH Paleobiology collection and digitization strategies at the Smithsonian Institution

Kathy Hollis
Collections Manager
NMNH Department of Paleobiology
hollisk@si.edu
Conceptualizing Paleobiology

You can’t manage what you don’t measure

- Estimated 40 million individual items
- Estimated 2 million Records
- 490 Collections
- 58 Collection Groups
History of digitization at NMNH Paleo

1960
SELGEM Developed and NMNH digitization began

1965
SELGEM Developed and NMNH digitization began

1970
First Paleo record entered

1979-1984
Congressional Inventory Project
150,000 of high-priority records entered by 1984
25 people on collections staff

1980

1990
2000
2010

Aug 1998
Paleo begins using a digital system to manage transactions

Nov 2001
Last Paleo records entered in SELGEM

Aug 2004
Paleo “go live” in EMu

Aug 2013
600,000th Paleo record entered
11 people on collections staff
At current rate, with current staff, with current workflows, NMNH Paleo Collection will be digitized in 400 years (3,500 per year).
At average digitization rate since 1970
NMNH Paleo Collection will be digitized in 100 years
(14,000 records per year).
At the Congressional Inventory digitization rate NMNH Paleo Collection will be digitized in 47 years (30,000 records per year)
History of digitization at NMNH Paleo

As of 2011, SI is requiring a systematic, priority-driven, centrally managed approach.
## Contemporary Digitization at SI

### Step 1. Define standards for record quality

<table>
<thead>
<tr>
<th>Emu Record Quality</th>
<th>Does not meet NMNH Paleo Standards and needs more data to be scientifically useful</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank 1</td>
<td>NOT DIGITIZED</td>
</tr>
<tr>
<td>Rank 2</td>
<td>One or more required fields are missing data</td>
</tr>
<tr>
<td>Rank 3</td>
<td>All required fields are complete</td>
</tr>
<tr>
<td>Rank 4</td>
<td>All Rank 3 plus if type, HVCI, or dinosaur then has an image</td>
</tr>
<tr>
<td>Rank 5</td>
<td>All Rank 4 plus properly georeferenced</td>
</tr>
</tbody>
</table>

Meets or exceeds NMNH Paleo Standards and are minimally scientifically useful

Fields are mapped in Emu to Darwin Core Fields and are auto populated
Contemporary Digitization at SI

Step 2. Gather Digitization Statistics

<table>
<thead>
<tr>
<th>Emu Record Quality</th>
<th>Total Emu Records as of 2013</th>
<th>Total Item Count (somewhat meaningless)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NOT DIGITIZED</td>
<td>1,280,578</td>
<td>36,508,960</td>
</tr>
<tr>
<td>Rank 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>One or more required fields are missing data</td>
<td>156,229</td>
<td>392,135</td>
</tr>
<tr>
<td>Rank 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All required fields are complete</td>
<td>443,773</td>
<td>7,185,593</td>
</tr>
<tr>
<td>Rank 4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All Rank 3 plus if type, HVCI, or dinosaur then has an image</td>
<td>8,277</td>
<td>20,775</td>
</tr>
<tr>
<td>Rank 5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All Rank 4 plus properly georeferenced</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1,888,857</td>
<td>44,107,463</td>
</tr>
</tbody>
</table>
Contemporary Digitization at SI

Step 3. Set Priorities and Goals

At the Smithsonian, priorities and goals are formalized in Director-approved 3-Year Digitization Plans and are managed through the Digitization Program Office.
Contemporary Digitization at SI

Step 3. Set Priorities and Goals

At the Smithsonian, priorities and goals are formalized in Director-approved 3-Year Digitization Plans and are managed through the Digitization Program Office.

A written digitization plan provides:
1. Unified vision for digitization activities
2. Digitization history
3. Alignment with Strategic Plan
4. Digitization scope, goals, and priorities
5. Capacity and resources
6. Collaborators and dependencies (e.g. funding sources)
7. Policies, standards, and legal restrictions (e.g. copyright)
8. Roles and responsibilities.
Step 3. Set Priorities and Goals - **Paleobiology’s 2015 goals:**

**Goal 1:** Create new specimen and locality records through single-record entry or migration of data sets; all will be at least to Rank 3 by 2015

**Goal 2:** Create new or migrate existing specimen images and attach them to specimen records, all will be at least to Rank 3 by 2015 *See Lightning Slide for details*

**Goal 3:** EMu maintenance, standardization, and data clean up in order to share data in online collections or biodiversity data portals by 2015
Contemporary Digitization at SI

Step 3. Set Priorities and Goals - **Paleobiology’s 2015 goals:**

**Goal 1:** Create new specimen and locality records through single-record entry or migration of data sets; all will be at least to Rank 3 by 2015.

See Lightning Slide for details.

**Goal 2:** Create new or migrate existing specimen images and attach them to specimen records; all will be at least to Rank 3 by 2015.

**Goal 3:** EMu maintenance, standardization, and data cleanup in order to share data in online collections or biodiversity data portals by 2015.
Step 4. Figure out logistics and workflows

At the Smithsonian, each digitization project should be managed through an approved Digital Asset Management Plan, which are filed at the Digitization Program Office.
Contemporary Digitization at SI

Step 4. Figure out logistics and workflows

A written Project Plan is for:
1. Stating the significance of the digital assets
2. How the project supports museum and Strategic goals
3. Stating the scope including type, volume, and owner of digital assets
4. Workflow
5. Stating audience and usage for digital assets
6. Policies and restrictions
7. Lifecycle management
8. Technical requirements
9. Roles and responsibilities
<table>
<thead>
<tr>
<th>Digitization Project</th>
<th>Records created by end of 2015</th>
<th>Percent of collection databased by 2015</th>
<th>Workflow</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bown-Rose Eocene Mammals</td>
<td>9,000</td>
<td>45%</td>
<td>Pull specimens, type data into Emu from specimen labels, verify data as needed</td>
</tr>
<tr>
<td>Ongoing Type specimen cataloging</td>
<td>Records created as needed; “filler” work</td>
<td>?%</td>
<td>Pull specimens, type data into Emu from specimen labels, verify data as needed</td>
</tr>
<tr>
<td>Ostracoda Biologic</td>
<td>2,000</td>
<td>100%</td>
<td>Pull specimens, type data into Emu from specimen labels, verify data as needed</td>
</tr>
<tr>
<td>All Exhibit specimens</td>
<td>2,000</td>
<td>100%</td>
<td>Create inventory labels for all specimens, update Excel file pre-mapped to Emu fields, migrate final data set to Emu.</td>
</tr>
<tr>
<td>Paleobotany (Cleared Leaves and Pz Plants)</td>
<td>6,000</td>
<td>44%</td>
<td>Pull specimens, type data into pre-mapped Emu migration template rom specimen labels, verify data as needed; Map Access database to Emu, standardize data where necessary, and migrate to EMu.</td>
</tr>
<tr>
<td>Paleobotany Localities</td>
<td>1,500</td>
<td>100%</td>
<td>Map Access database to Emu, standardize data where necessary, and migrate to EMu.</td>
</tr>
<tr>
<td>Green River plants and insects</td>
<td>30,000</td>
<td>100%</td>
<td>Map Filemaker database to Emu, standardize data where necessary, and migrate to Emu.</td>
</tr>
</tbody>
</table>
NMNH Paleo tips for undertaking digitization projects

1. Collect data on where you are starting.

2. Write plan and priorities for what you want to digitize in the next few years.

3. Define written standards for record content and quality, especially if workflow varies among digitization projects. Standards should meet interoperability requirements.

4. Centralized, stable storage during data and image processing is essential. No free-standing data on external hard drives!
NMNH Paleo tips for undertaking digitization projects

1. Collect data on where you are starting.

2. Write plan and priorities for what you want to digitize in the next few years.

3. Define written standards for record content and quality, especially if workflow varies among digitization projects. Standards should meet interoperability requirements.

4. Centralized, stable storage during data and image processing is essential. No free-standing data on external hard drives!

Next steps for NMNH Paleo digitization:

1. Learn more about efficient, time-saving workflows and put them in place.

2. Educate department on digitization and data sharing so it is a universal priority.

3. Start a georeferencing initiative.

4. Set our data free!

hollisk@si.edu