Emerging Pathways for the Transcription of Scientific Field Notes

CLIR DHC Symposium
October 12, 2022 | Baltimore, MD
Field Notes: Invaluable Primary Resources

Smithsonian Institution Archives, SIA RU007203, Created by Murdoch, John, 1852-1925, "Record of collections made by the Point Barrow Polar Expedition, 1881-1883", SIA2015-001147
Our Panel

**Riccardo Ferrante**  
Associate Director of Information Systems, Digital Lifecycles and User Experience  
Smithsonian Libraries and Archives

**Sonoe Nakasone**  
Data Specialist  
American Women's History Initiative and Smithsonian American Art Museum

**Katie Mika**  
Data Services Librarian  
Harvard Library and Institute for Quantitative Social Science (IQSS)
A good start
Making scientific field notes accessible
with The Field Book Project
By Riccardo Ferrante, he/him
Building a road of discovery

• First things first - where are they?

  • In 2009, the CLIR Cataloging Hidden Collections program funded the Smithsonian’s Exposing Biodiversity Fieldbooks and Original Expedition Journals with two objectives

    • Catalog an estimated 6,000 fieldbooks and journals whose only online presence, if any, was as a line in an online archival finding aid.

    • Build and launch an online catalog of the field books making it accessible to researchers worldwide.

    • Utilize standards such as MODS and EAC-Cpf to facilitate data sharing with biodiversity and natural history organizations, libraries, archives and aggregators.
Consequences of a good thing

• Along the way

  • Item-level cataloging revealed a **pressing need for conservation** and preservation treatment. Working directly with decades- and century-old field journals and notes, project catalogers were trained to make preliminary preservation assessments and alert conservators when further handling of a field book posed such risk to its condition.

  • “**Have you digitized this?**” followed soon after by “**Do you have a transcript of this?**”
Pressing forward

• Preservation and Conservation
  • The National Park Service’s Save America’s Treasures grant program allowed us to stabilize at-risk field books.

• Digitization
  • Small internal grants funded internships to continue small project-based digitization.

• Transcription
  • An inaugural contributor to the Smithsonian’s crowd-sourced transcription platform provided a way to respond despite lack of funds
Pressing forward (more)

- Preservation and Conservation
- Digitization
- Rapid-capture digitization and more cataloging
  - A large grant from the Arcadia Foundation enabled the cataloging of over 3,000 additional scientific biodiversity-related field books, growing the catalog by 50%, and enabling us to implement rapid capture digitization of these materials.
- Access via more platforms, at a deeper level
  - Digitized field books were added to the Field Book Registry, the Biodiversity Heritage Library, and the Digital Public Library of America.
  - Crowd-sourced transcriptions create a full-text searching option on those platforms, also Google.
Even better together

In 2015, CLIR Launches the *Digitizing Hidden Collections* program.

Ten BHL partners came together to further expand the scientific field notes collections in the Biodiversity Heritage Library.

**Biodiversity Heritage Library Field Notes Project**

- Smithsonian Institution
- Missouri Botanical Garden, Peter H. Raven Library;
- American Museum of Natural History;
- Yale Peabody Museum;
- Harvard University, Herbaria Botany Libraries;
- Harvard Museum of Comparative Zoology, Ernst Mayr Library;
- University of California, Berkeley Museum of Vertebrate Zoology;
- New York Botanical Garden, LuEsther T. Mertz Library;
- The Field Museum;
- Internet Archive
It’s about access, right?

• The access-based empowerment we expected
  • From cataloging
    • Taxonomic names, geolocations, subjects, authors
  • From digitization
    • Viewing the item
    • If printed, OCR is possible for full text and specialized indexing

• Increased use and downloads globally
Empowerment surprises

• Transcription impact on **handwritten** field notes
  - Paleography is a dying art, but far from dead yet. Thousands of volunteers worked together, producing high quality transcripts when there were no funds to be found. (thank you, Internet crowd!)
  - Full text searching became possible, including searching across handwritten and printed material, i.e., primary and secondary sources together or apart.
  - Taxonomy links between historic and current names, both scientific and common
  - Hidden or misrepresented collectors
Empowerment surprises

Transcription revealed hidden or misrepresented collectors, especially women

• Example: Joseph Nelson Rose's 1851 catalog of specimens documented over 25 women contributors

• Meet Mrs. Katherine Ross Gaillard. Misattributed in several catalogs as “D. D. Gaillard,” these entries can now be properly attributed.
Untapped potential

• Linked data
Untapped potential

- Extracting data sets from handwritten sources

F.W. True and Daniel W. Prentiss, Jr. “Maine, 1897.” Pages 6 and 7 of this field catalog.
Thank you.
Legacy Data to Linked Data
How Linked Data Could Change Field Book Metadata
By Sonoe Nakasone, she/they
Vision of 2009 Field Book Project

• Repository of metadata about field books

• Detailed item level descriptions of field books

• Save time and money on travel for research
'Linked Data' = 2006 | Our project = 2009

• Field Book Project used XML schemas: MODS, NCD, and EAC-CPF

• 'Linked data' was around, but new and unfamiliar to many in libraries
If we did it LD

Pearl and Hermes Reef --> Honolulu --> Hawai'i --> United States

Location subjects in field book record
Pearl and Hermes Reef
Hawai'i
United States

EVERY TIME, in all 49 items matching "Pearl and Hermes Reef"

Works for queries: 1) Show me all "Pearl and Hermes Reef" field notes; 2) all "Hawai'i" field notes; and 3) all "United States" field notes
If we did it LD

Pearl and Hermes Reef --> Honolulu --> Hawaiʻi --> United States

As triples...
"Field book X" "depicts location" "Pearl and Hermes Reef". (EVERY TIME)

"Pearl and Hermes Reef" "is part of" "Hawaiʻi". (ONCE)
"Hawaiʻi" "is part of" "United States". (ONCE)

Works for queries: 1) Show me all "Pearl and Hermes Reef" field notes;
2) all "Hawaiʻi" and its subsidiaries field notes; and 3) all "United States" and its subsidiaries field notes
Example components of a Linked Data stack

- **Search / display GUI**
  - Using SPARQL

- **Metadata GUI**
  - Optional to simplify creating new statements

- **SPARQL endpoint**
  - Querying service for RDF

- **Search index**
  - Optional to simplify search

- **Triplestore**
  - Database of RDF Statements

---

Smithsonian Libraries and Archives
Biodiversity Heritage Library
HARVARD LIBRARY
Wikibase, Wikidata, or similar

• Ready to use linked data infrastructure


• Connectivity to other data on the web and search engine users
Title: Pearl and Hermes Reef, September 1966, Lewis
Creator: LEWIS, T. JAMES
Repository location: Smithsonian Institution Archives
Places: PEARL AND HERMES REEF
HAWAII
UNITED STATES
Start date: 1966-09-25
End date: 1966-09-27
Topic: PACIFIC OCEAN BIOLOGICAL SURVEY PROGRAM
Ornithology
Accession Number: SIA RU000245

Looking ahead toward LD for field books

• Field book metadata is already structured data

• Some data shouldn't be open

• Shared data model to aide interoperability
Thank you

Nakasones@si.edu
Extracting Datasets from Digital Collections

Katie Mika
Data Services Librarian
Harvard Library & Institute for Quantitative Social Science
Kinds of data

How to extract data & datasets from items

Collections as Data as a distinct service model
Kinds of Data
“Data refers to entities used as evidence of phenomena for the purposes of research or scholarship.”

“Data are not pure or natural objects with an essence of their own. They exist in a context, taking on meaning from that context and the perspective of the beholder.”

LAPETROSIJA CRISTATA
Blood-splotted Lapetrosia
Bulb of South Africa

Family: LILIACEAE
Subfamily: LILIOIDEAE

Lapetrosia is a genus of some forty-five species, mostly African, inhabiting the dry Cape region, and one of the most beautiful garden plants from that country, whose name is standard, and known as Chrysanthemum. From the same region and area produced this fragrant species of greenhouse, the belladonna lily, and each other better known as Heliconia, Anthurium, or umbrella lilies. Pedalium, Convallaria, Muscari and Draca, all brilliantly colored. Like the gladiolus, Lapetrosia has a more delicate flower), and may be propagated in much the same way, as it is raised in the same manner as the popular belladonna. The species can be grown on open ground or under cover. Several from South Africa are probably hardy and are among the most beautiful and effective flowers.

Pedals are supported on thin, long, green stems; the leaves are large, oval, and greenish, and resemble those of the calla lilies. The flowers are produced in long, narrow, spike-like clusters, and the petals are large, pink, and reflexed. The plant is hardy, and is propagated by division in the autumn. The flowers are produced during the summer and autumn months.

Pedalium munitum. A variety of Pedalium munitum, with large, green, and reflexed leaves, and large, pink, and reflexed, bell-shaped flowers. The flowers are produced during the summer and autumn months.

Pedalium munitum. A variety of Pedalium munitum, with large, green, and reflexed leaves, and large, pink, and reflexed, bell-shaped flowers. The flowers are produced during the summer and autumn months.

Lapetrosia is a genus of some forty-five species, mostly African, inhabiting the dry Cape region, and one of the most beautiful garden plants from that country, whose name is standard, and known as Chrysanthemum. From the same region and area produced this fragrant species of greenhouse, the belladonna lily, and each other better known as Heliconia, Anthurium, or umbrella lilies. Pedalium, Convallaria, Muscari and Draca, all brilliantly colored. Like the gladiolus, Lapetrosia has a more delicate flower), and may be propagated in much the same way, as it is raised in the same manner as the popular belladonna. The species can be grown on open ground or under cover. Several from South Africa are probably hardy and are among the most beautiful and effective flowers.


<table>
<thead>
<tr>
<th>Date</th>
<th>Event</th>
</tr>
</thead>
</table>

William H. Dall - Field Notes, 1880 (1 of 2) | Smithsonian Digital Volunteers.” Accessed October 7, 2022. [Transcription](https://transcription.si.edu/project/6980/)
Show me all "Pearl and Hermes Reef" field notes; all "Hawai'i" and its subsidiaries field notes; and all "United States" and its subsidiaries field notes

Show me all “Puffinus pacificus” field notes; where ”date” is “July”

How to extract data and datasets from items
<table>
<thead>
<tr>
<th>Field</th>
<th>Order</th>
<th>Species</th>
<th>SVH Bole</th>
<th>Length</th>
<th>Position</th>
<th>Genus</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>59</td>
<td>M. robusta</td>
<td>6</td>
<td>235</td>
<td>m. robusta</td>
<td>M. robusta</td>
</tr>
<tr>
<td>2</td>
<td>60</td>
<td>M. robusta</td>
<td>7</td>
<td>235</td>
<td>m. robusta</td>
<td>M. robusta</td>
</tr>
<tr>
<td>3</td>
<td>61</td>
<td>M. robusta</td>
<td>8</td>
<td>235</td>
<td>m. robusta</td>
<td>M. robusta</td>
</tr>
<tr>
<td>4</td>
<td>62</td>
<td>M. robusta</td>
<td>9</td>
<td>235</td>
<td>m. robusta</td>
<td>M. robusta</td>
</tr>
<tr>
<td>5</td>
<td>63</td>
<td>M. robusta</td>
<td>10</td>
<td>235</td>
<td>m. robusta</td>
<td>M. robusta</td>
</tr>
<tr>
<td>6</td>
<td>64</td>
<td>M. robusta</td>
<td>11</td>
<td>235</td>
<td>m. robusta</td>
<td>M. robusta</td>
</tr>
<tr>
<td>7</td>
<td>65</td>
<td>M. robusta</td>
<td>12</td>
<td>235</td>
<td>m. robusta</td>
<td>M. robusta</td>
</tr>
<tr>
<td>8</td>
<td>66</td>
<td>M. robusta</td>
<td>13</td>
<td>235</td>
<td>m. robusta</td>
<td>M. robusta</td>
</tr>
</tbody>
</table>

**Notes:**
- SVH Bole: S∨H Bole
- Length: Length
- Position: Position
- Genus: Genus
<table>
<thead>
<tr>
<th>Field Number</th>
<th>SPGCE35</th>
<th>SLK</th>
<th>DLN</th>
<th>3004</th>
<th>LOCATION</th>
<th>FAM</th>
<th>REPRODUCTIVE</th>
<th>BIOLOGY</th>
<th>SPECIES</th>
<th>STATUS</th>
</tr>
</thead>
<tbody>
<tr>
<td>2209</td>
<td>2209</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2320</td>
<td>2320</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2330</td>
<td>2330</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note: The image contains a table with columns for Field Number, SPGCE35, SLK, DLN, 3004, LOCATION, FAM, REPRODUCTIVE, BIOLOGY, SPECIES, and STATUS. The table contents are not fully visible due to the resolution of the image.*
<table>
<thead>
<tr>
<th>Column 1</th>
<th>Column 2</th>
<th>Column 3</th>
<th>Column 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value 1</td>
<td>Value 2</td>
<td>Value 3</td>
<td>Value 4</td>
</tr>
<tr>
<td>Value 5</td>
<td>Value 6</td>
<td>Value 7</td>
<td>Value 8</td>
</tr>
<tr>
<td>Value 9</td>
<td>Value 10</td>
<td>Value 11</td>
<td>Value 12</td>
</tr>
<tr>
<td>Value 13</td>
<td>Value 14</td>
<td>Value 15</td>
<td>Value 16</td>
</tr>
</tbody>
</table>

Haber 872
Creating data tables from images of manuscript field notes

This script uses the bounding box coordinates to generate a pandas DataFrame

Bounding box notes:
Bounding boxes are made up of 6 values: x1, y1, x2, y2, x3, y3. The values are in the format of (x1,y1, x2,y2, x3,y3). This information helps to create the dictionaries of x, y values used throughout the script.

1. Setup

    - Import libraries and data
    - Setup vars immediately
    - Remove external libraries

In [151]:
# Import libraries and data
import pandas
import numpy as np
import matplotlib
import seaborn
import matplotlib.pyplot as plt

In [152]:
# Set of lines
for i in range(10):
    ...:            line = random.choice([0, 1])

In [153]:
# Add all x, y properties
for k in range(10):
    ...:        x = random.randint(0, 10)

In [154]:
# Print all y properties
for k in range(10):
    ...:        y = random.randint(0, 10)

<table>
<thead>
<tr>
<th>cell</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>2269</td>
<td>Phleothorax rubricinctus</td>
<td>#NAME?</td>
<td>6-July</td>
<td>180.05 N 160.40 W</td>
<td>670°</td>
<td>Outdoor</td>
<td>L. 5w2</td>
<td>Notes</td>
<td>Color</td>
<td>Notes</td>
<td>Huber 838</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>61</td>
<td>Phleothorax rubricinctus</td>
<td>7-July</td>
<td>150.1 N 130.15 W</td>
<td>47.37</td>
<td>-</td>
<td>L. ovum</td>
<td>mm</td>
<td></td>
<td></td>
<td></td>
<td>Chandler 0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>61</td>
<td>Phleothorax rubricinctus</td>
<td>7-July</td>
<td>160.03 N 1630.25 W</td>
<td>82.8</td>
<td>-</td>
<td>ES8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>62</td>
<td>P. hypolca nigromelas</td>
<td>0-July</td>
<td>110.0 N 100.22 W</td>
<td>very light</td>
<td>Daley 8X 44-1</td>
<td>21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>63</td>
<td>P. hypolca nigromelas</td>
<td>0-July</td>
<td>119.34 N 160.27 W</td>
<td>134.7</td>
<td>Light - mid</td>
<td>Black</td>
<td>2X1</td>
<td>mm</td>
<td></td>
<td></td>
<td>Huber 837</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>64</td>
<td>SterNA tenuata</td>
<td>080 000 1100 000 0</td>
<td>1154</td>
<td>JonArS Not found</td>
<td>color Notes</td>
<td>PEARSON 23</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>65</td>
<td>Puffinus pacificus</td>
<td>300</td>
<td>Light - mod</td>
<td>Rodent 7X 2-3m</td>
<td>Huber 839</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>70</td>
<td>Puffinus pacificus</td>
<td>300</td>
<td>Light</td>
<td>Rodent 7X 2-3m</td>
<td>Huber 839</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>68</td>
<td>Streunornis antarctica</td>
<td>000 000 000 000 0</td>
<td>850</td>
<td>Very light</td>
<td>4X4/10 X 25 X</td>
<td>60</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>70</td>
<td>Streunornis antarctica</td>
<td>000 000 000 000 0</td>
<td>850</td>
<td>Very light</td>
<td>4X4/10 X 25 X</td>
<td>60</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>71</td>
<td>Streunornis antarctica</td>
<td>000 000 000 000 0</td>
<td>850</td>
<td>Very light</td>
<td>4X4/10 X 25 X</td>
<td>60</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>72</td>
<td>Streunornis antarctica</td>
<td>000 000 000 000 0</td>
<td>850</td>
<td>Very light</td>
<td>4X4/10 X 25 X</td>
<td>60</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>73</td>
<td>Streunornis antarctica</td>
<td>000 000 000 000 0</td>
<td>850</td>
<td>Very light</td>
<td>4X4/10 X 25 X</td>
<td>60</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>74</td>
<td>Streunornis antarctica</td>
<td>000 000 000 000 0</td>
<td>850</td>
<td>Very light</td>
<td>4X4/10 X 25 X</td>
<td>60</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>75</td>
<td>Streunornis antarctica</td>
<td>000 000 000 000 0</td>
<td>850</td>
<td>Very light</td>
<td>4X4/10 X 25 X</td>
<td>60</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>76</td>
<td>Streunornis antarctica</td>
<td>000 000 000 000 0</td>
<td>850</td>
<td>Very light</td>
<td>4X4/10 X 25 X</td>
<td>60</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>77</td>
<td>Streunornis antarctica</td>
<td>000 000 000 000 0</td>
<td>850</td>
<td>Very light</td>
<td>4X4/10 X 25 X</td>
<td>60</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>78</td>
<td>Streunornis antarctica</td>
<td>000 000 000 000 0</td>
<td>850</td>
<td>Very light</td>
<td>4X4/10 X 25 X</td>
<td>60</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>79</td>
<td>Streunornis antarctica</td>
<td>000 000 000 000 0</td>
<td>850</td>
<td>Very light</td>
<td>4X4/10 X 25 X</td>
<td>60</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>80</td>
<td>Streunornis antarctica</td>
<td>000 000 000 000 0</td>
<td>850</td>
<td>Very light</td>
<td>4X4/10 X 25 X</td>
<td>60</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>81</td>
<td>Streunornis antarctica</td>
<td>000 000 000 000 0</td>
<td>850</td>
<td>Very light</td>
<td>4X4/10 X 25 X</td>
<td>60</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[Link to the original dataset](https://github.com/kmika11/bhl_unlocking_datatables/blob/main/sea19631966196800natib_0004.csv)
What we learned

• Can work for optimal items
• Contribute datasets to Open Data landscape
• Validation is iterative
• “Low hanging fruit”
Collections as Data as a distinct service model
<table>
<thead>
<tr>
<th>Date</th>
<th>Specimen</th>
<th>Location</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>11/3/13</td>
<td></td>
<td>Cape York, Queensland, Australia</td>
<td></td>
</tr>
<tr>
<td>11/4/13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11/5/13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11/6/13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11/7/13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11/8/13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11/9/13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11/10/13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11/11/13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11/12/13</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Thanks!