Authors & Affiliations

Katie Mika, Data Services Librarian, Harvard Library

JJ Dearborn, Data Manager, Biodiversity Heritage Library / Smithsonian Libraries and Archives

Extracting Expedition Log Data Found in the Biodiversity Heritage Library

The Biodiversity Heritage Library (BHL) has amassed a large collection of expedition logs and archival field notes. The data contained in these materials are non-standard, unreadable by machines, and unvetted by humans.

We Need Feedback!

If you are a computational researcher, bioinformatician, museum professional, or environmental policy analyst interested in making use of historic species data, please:

→ Click here to answer a brief survey

Introduction

Species occurrence data contained in expedition logs can reveal snapshots of past ecological states and illuminate human impacts on species habitats. Extracting this data from BHL's historic texts will increase humanity's understanding of environmental change through time at hyperlocal and global scales.

Objective

BHL staff would like to deposit invaluable historic species occurrence data with big data biodiversity aggregators like the Global Biodiversity Information Facility (GBIF) and information brokers like Wikidata. Once deposited, this data could then be leveraged by climate change researchers, developers of global species monitoring platforms, and environmental policy makers.

Methodology

Piloted the application of Handprint, a Handwriting Text Recognition (HTR) algorithm, to images of analog field notes and used bounding box outputs to sort text into a data table.

Results

Handwriting Text Recognition (HTR) is still unable to reliably transform handwriting into machine readable text, but it is improving. It is now possible to sort text output into tables using bounding box coordinates.

Anaylsis

Handprint, an HTR program developed at CalTech Library, automatically groups together identified text items into *lines*, which roughly correspond to *cells* in a data table. We were then able to identify groups of boxes that overlapped along the x- and y-axes to create columns and rows.

Because the handwritten table was neatly spaced, the pipeline was largely successful. This process will be less effective for messier data tables.

Blue-faced boobies, Pacific Ocean, November 11, 1968.

Lesser frigate-bird egg, Jarvis Island, circa 1961-1973.

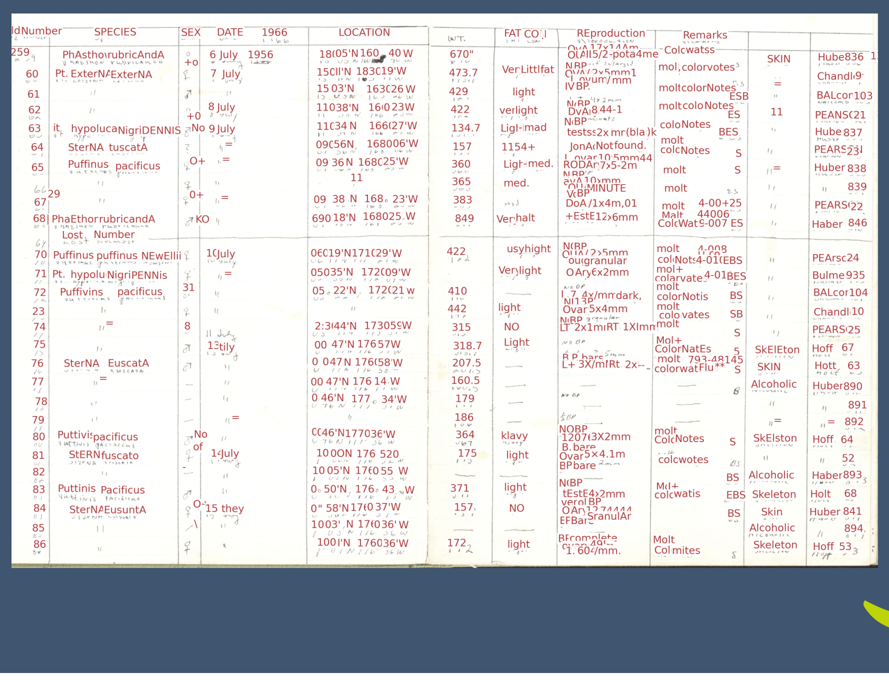
Fairy tern, Pacific Ocean, circa 1961-1973.

The Sample Set

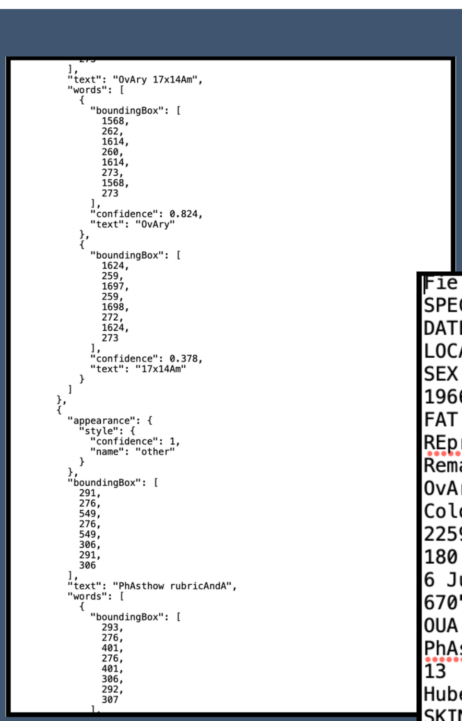
Birds observed, banded, and collected as part of the Pacific Ocean Biological Survey Program (POBSP), 1961-1973. The Program deployed over 40 Smithsonian Institution employees to conduct biological surveys of plants and animals that occurred on the islands and atolls. A major focus was determining migration, distribution, and populations of seabirds.

BEFORE

handprint_output.**png**



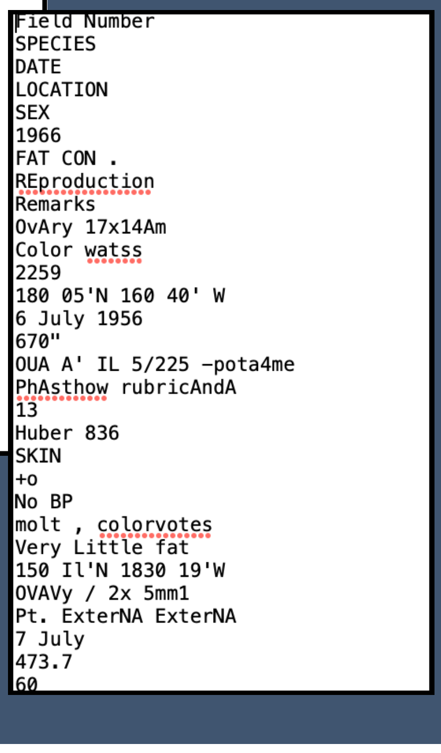
Field Number
SPECIES
SEX
DATE
1966
LOCATION
WT.
FAT CON
REproduction
Remarks
SKIN



```
      ],
      "text": "OvAry 17x14Am",
      "words": [
        {
          "boundingBox": [
            1568,
            262,
            1614,
            260,
            1614,
            273,
            1568,
            273
          ],
          "confidence": 0.824,
          "text": "OvAry"
        },
        {
          "boundingBox": [
            1624,
            259,
            1697,
            259,
            1698,
            272,
            1624,
            273
          ],
          "confidence": 0.378,
          "text": "17x14Am"
        }
      ]
    },
    {
      "appearance": {
        "style": {
          "confidence": 1,
          "name": "other"
        }
      },
      "boundingBox": [
        291,
        276,
        549,
        276,
        549,
        306,
        291,
        306
      ],
      "text": "PhAsthow rubricAndA",
      "words": [
        {
          "boundingBox": [
            293,
            276,
            401,
            276,
            401,
            306,
            292,
            307
```

handprint_output.**json**

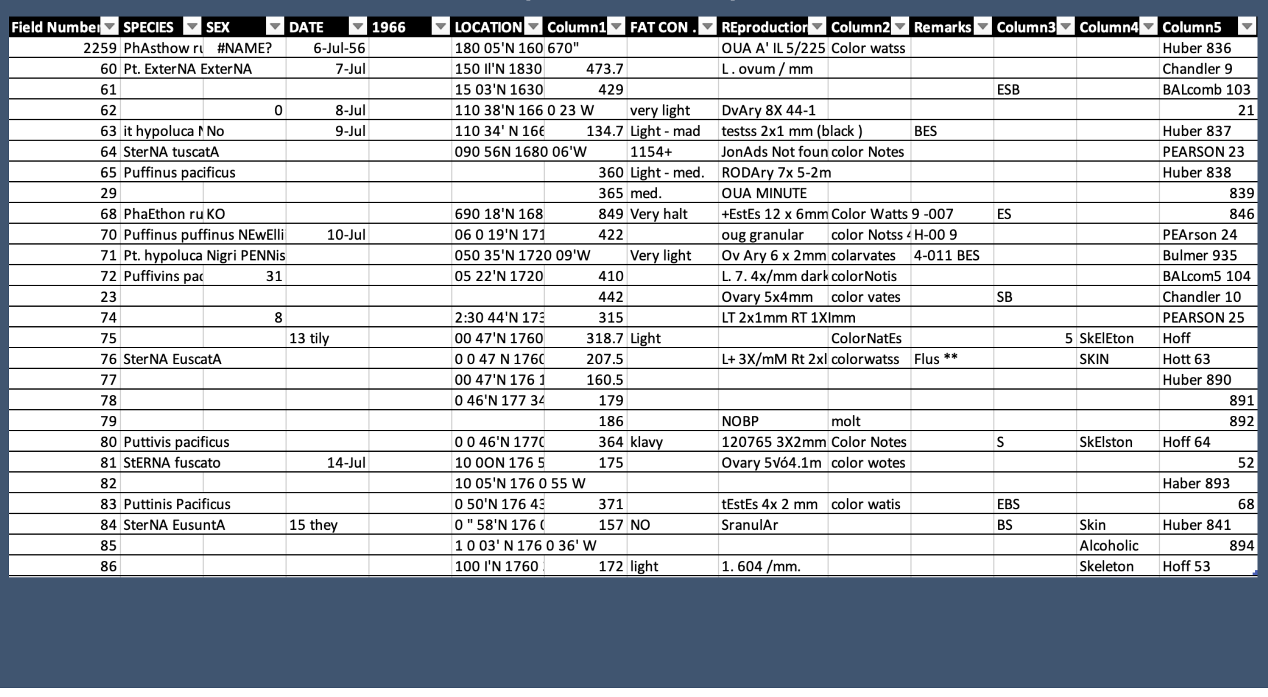
handprint_output.**txt**



```
Field Number
SPECIES
DATE
LOCATION
SEX
1966
FAT CON .
REproduction
Remarks
OvAry 17x14Am
Color watss
2259
180 05'N 160 40' W
6 July 1956
670"
OUA A' IL 5/225 -pota4me
PhAsthow rubricAndA
13
Huber 836
SKIN
+o
No BP
molt , colorvotes
Very Little fat
150 Il'N 1830 19'W
OVAVy / 2x 5mm1
Pt. ExterNA ExterNA
7 July
473.7
60
```

AFTER

handprint_output.**csv**



Field Number	SPECIES	SEX	DATE	1966	LOCATION	Column1	FAT CON .	REproduction	Column2	Remarks	Column3	Column4	Column5
2259	PhAsthow ru	#NAME?	6-Jul-56		180 05'N 160	670"		OUA A' IL 5/225	Color watss				Huber 836
60	Pt. ExterNA ExterNA		7-Jul		150 Il'N 1830	473.7		L . ovum / mm					Chandler 9
61					15 03'N 1630	429					ESB		BALcomb 103
62		0	8-Jul		110 38'N 166 0 23 W		very light	DvAry 8X 44-1					21
63	it hypoluca N	No	9-Jul		110 34' N 166	134.7	Light - mad	testss 2x1 mm (black)		BES			Huber 837
64	SterNA tuscatA				090 56N 1680 06'W		1154+	JonAds Not foun	color Notes				PEARSON 23
65	Puffinus pacificus					360	Light - med.	RODAry 7x 5-2m					Huber 838
29						365	med.	OUA MINUTE					839
68	PhaEthon ru	KO			690 18'N 168	849	Very halt	+EstEs 12 x 6mm	Color Watts 9 -007		ES		846
70	Puffinus puffinus NEwElli		10-Jul		06 0 19'N 171	422		oug granular	color Notss 4	H-00 9			PEArson 24
71	Pt. hypoluca Nigri PENNis				050 35'N 1720 09'W		Very light	Ov Ary 6 x 2mm	colarvates	4-011 BES			Bulmer 935
72	Puffivins pac	31			05 22'N 1720	410		L. 7. 4x/mm dark	colorNotis				BALcom5 104
23						442		Ovary 5x4mm	color vates		SB		Chandler 10
74		8			2:30 44'N 173	315		LT 2x1mm RT 1XImm					PEARSON 25
75			13 tily		00 47'N 1760	318.7	Light		ColorNatEs		5	SkElEton	Hoff
76	SterNA EuscatA				0 0 47 N 1760	207.5		L+ 3X/mM Rt 2xl	colorwatss	Flus **		SKIN	Hott 63
77					00 47'N 176 1	160.5							Huber 890
78					0 46'N 177 34	179							891
79						186		NOBP	molt				892
80	Puttivis pacificus				0 0 46'N 1770	364	klavy	120765 3X2mm	Color Notes		S	SkElston	Hoff 64
81	StERNA fuscato		14-Jul		10 0ON 176 5	175		Ovary 5v64.1m	color wotes				52
82					10 05'N 176 0 55 W								Haber 893
83	Puttinis Pacificus				0 50'N 176 43	371		tEstEs 4x 2 mm	color watis		EBS		68
84	SterNA EusuntA		15 they		0 " 58'N 176 0	157	NO	SranulAr			BS	Skin	Huber 841
85					1 0 03' N 176 0 36' W							Alcoholic	894
86					100 I'N 1760	172	light	1. 604 /mm.				Skeleton	Hoff 53

Conclusion

Once images of occurrence records are transformed into machine readable data, species names can be validated, identifiers can be added, datasets can be deposited to biodiversity data aggregators, and catalog records of literature in BHL can be better connected to specimen records in natural history museums.

Data & Materials

Data: National Museum of Natural History (U.S.) Pacific Ocean Biological Survey Program (1966). At-sea, 1963-1966, 1968, part 3 : July - August 1966 . 1966. https://doi.org/10.5962/bhl.title.148243.

Images: Smithsonian Institution. (2022, May 12). Pacific Ocean Biological Survey Program. Flickr. Retrieved May 1, 2022, from https://www.flickr.com/photos/smithsonian/albums/72157627185361301

Handprint Handwriting Text Recognition: https://github.com/caltechlibrary/handprint

Transformation code available at: https://github.com/kmika11/bhl_unlocking_datatables

Poster Presented at SPNHC 2022 | Liberating Natural History Collections Data in Biodiversity Literature