https://doi.org/10.3897/biss.2.26222

Smithsonian Institution

NATIONAL MUSEUM of NATURAL HISTORY

Digitizing EPICC Data: Trials and tribulations in translating 100 year old data

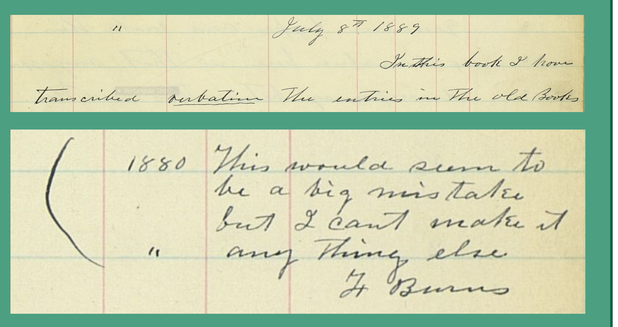
Holly Little, Anna Leary, Alexandra Cano, Adam T. Mansur

The Challenge

The Smithsonian National Museum of Natural History (NMNH) Department of Paleobiology recently completed the first segment of a mass digitization project in support of the Eastern Pacific Invertebrate Communities of the Cenozoic (EPICC) thematic collections network. In collaboration with the Smithsonian Institution Digitization Project Office (DPO), imaging and label transcription was completed for a portion of the Cenozoic Mollusca Collection. Once the labels were transcribed further processing was required to clean and enhance that specimen data. We sought to ensure high quality data for this project through:

1. the development of clear guidelines for documentation and treatment of specific data points;
2. updating records to match current taxonomic, lithostratigraphic, and chronostratigraphic information; and
3. create iterative workflows to maintain extensibility and to capture uncertainty in the data.

A significant challenge for any large collections digitization project is transcribing and cleaning analog information from specimen labels. The following considerations, methods, and tools enabled us to face these challenges of translating analog collections data of over a hundred years old into high quality records following modern standards for biodiversity information.

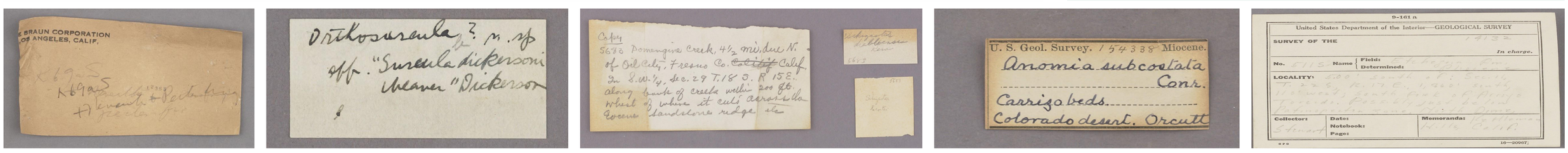


These challenges are not new! During this work we discovered ledger notes written by previous collections staff ~100 years ago that still ring true today

NMNH Paleo EPICC Stats

- 44,137 specimen lots imaged and databased
- 36,007 images of standard orientations
- 27,246 labels images and transcribed
- >220,000 specimens captured

EPICC is a partnership of nine natural history museums, united to digitize marine invertebrate fossils found in the eastern Pacific. The bivalves, gastropods, crabs, echinoids and other invertebrates found in our collections span the last 66 million years of Earth's history. During the course of the grant we will make 1.6 million specimen records available online through digital data and photographs. We are funded through the National Science Foundation's Advancing Digitization of Biological Collections program and affiliated with Integrated Digitized Biocollections (iDigBio). https://epicc.berkeley.edu/



1. Image Capture

With transcription, original data is processed into a standardized record. Imaging all labels creates a record of that original data. Metadata recorded in each specimen and label image was used to create skeleton records in the collections information system (EMu) and to initiate label transcription.

filename = catalog number captured from barcode
iptc:instructions = drawer location captured from drawer barcode
iptc:jobid = LocalityID in standardized format, found on label or specimen
(This process required creative use of iptc fields because the data did not map cleanly and is used for workflow only)

Prefix	Catalog	Event/Site: (Object D...	Current Locations/Current Locations: (Current Locations)/Location: (L...
PAL	685907	: : USGS LOC : 190...	NHB - East Wing - Basement - EB1B - Row 08 - Column 05 - Lower C...
PAL	684122		NHB - East Wing - Basement - EB1B - Row 08 - Column 02 - Lower C...
PAL	683548		NHB - East Wing - Basement - EB1B - Row 08 - Column 02 - Upper C...
PAL	681749	: : USGS LOC : 215...	NHB - East Wing - Basement - EB1B - Row 09 - Column 04 - Lower C...
PAL	684287	: : USGS LOC : 407...	NHB - East Wing - Basement - EB1B - Row 08 - Column 04 - Upper C...
PAL	684105	: : USGS LOC : 177...	NHB - East Wing - Basement - EB1B - Row 08 - Column 02 - Lower C...
PAL	684929	: : USGS LOC : 409...	NHB - East Wing - Basement - EB1B - Row 08 - Column 03 - Lower C...
PAL	685112	: : USGS LOC : 410...	NHB - East Wing - Basement - EB1B - Row 08 - Column 03 - Lower C...

2. Label Transcription

Transcription was completed in weekly batches creating a total of 21 datasets. It was important to create a workflow that would establish clear transcription guidelines to ensure consistency across the datasets and to account for the variety in data quality, quantity, and clarity as seen in the above label images.

Transcription Fields

1	Filename	Pulled from image file
2	Taxonomy	verbatim taxonomy string
3	Collector	
4	Geography	town, county, state
5	Location Description	
6	Additional Information	Any remaining information, labeled if possible (e.g. "Date: Aug 30-1917")

3. Merging and Parsing with Python

We developed a Python script to merge the skeleton records and transcription records and to parse any identifiable data from 'Additional Information' into more specific fields.

A benefit to creating the skeleton record from the image capture and then merging with transcription is that even specimens without labels will gain data.

Filename	CatNumb	LocSiteNu	LocSiteSta	BioEventS	LocBarcod	AdmDate	Taxonomy	Collector	Geography	Location Description	Geologic Ag	Addition
USNM_PAL_00695	695315	USGS LOC	8154	12071680	R10C10LD	3/5/2018	Pecten hemphil	Wagner	California	Camulos Quadrangle, West side Brown Canyo		
USNM_PAL_00695	695316	USGS LOC	8151	12071763	R10C10LD	3/5/2018	Dendraster paci	K	California	Camulos Quadrangle, In canyon between 2005		
USNM_PAL_00695	695317	USGS LOC	8153	12071650	R10C10LD	3/5/2018	Pecten cf. hemp	Wagner	California	Camulos Quadrangle, South slope of hill west		
USNM_PAL_00695	695318	USGS LOC	8152	12071810	R10C10LD	3/5/2018	Dendraster dieg	K	California	Camulos Quadrangle, In N.W. corner 8152 = 81		
USNM_PAL_00695	695319	USGS LOC	8141	12071635	R10C10LD	3/5/2018	Pecten (hinnites*) giganteus |	Phacoides annulatus | Olivella biplicata | Conus californic			814
USNM_PAL_00695	695320	USGS LOC	8156	12071842	R10C10LD	3/5/2018	Dend. Drezvensi	K	California	Camulos Quadrangle, North of Santa	8156=814	
USNM_PAL_00695	695321	USGS LOC	8157	12071712	R10C10LD	3/5/2018	Dendraster dieg	K	California	Santa Susana* Quad, on prominent hill about		
USNM_PAL_00695	695323	USGS LOC	18209	12071555	R10C10LD	3/5/2018	Prototthaca ? Cf.	Winterer, Wo	California	Ventura Basin	Pliocene	
USNM_PAL_00695	695324	USGS LOC	18209	12071555	R10C10LD	3/5/2018	Tellina cf. idae Dall		California	Ventura Basin	Pliocene	

Collector	Date (Parsed)	Ident By (I	Formatior	Locality (F	Memoran	No (Parse	Notebook	Pages (Pai	Field (Pa
Kew	Aug 30 - 1917				Near base	50	1	4	
Kew	Oct. 3, 1917				Fernandir	47			
Kew	30-Aug-17				Zone A of	49	1	4	
Kew	25-Oct-17				San Diego	48	2		
Kew					Pernauldc	52			
Kew	Aug 30, '17				! In a zone	53			

4. OpenRefine: Standardizing and Enhancing the Data

Each dataset is processed in OpenRefine, a powerful, open source data wrangling tool. With this tool we can complete data standardization and enhancement efficiently across these large datasets.

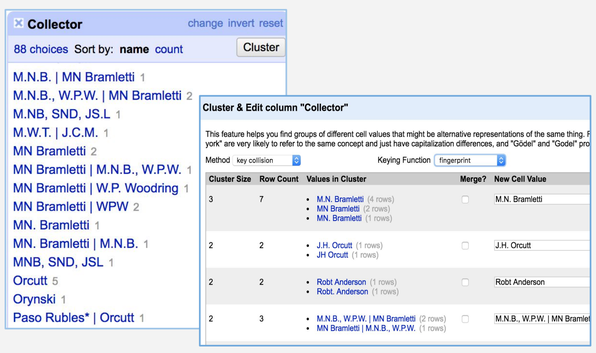
Standardization: Data from transcribed labels can vary in spelling, organization, and information. Using tools within OpenRefine such as clustering and faceting, patterns emerge enabling the cleaning and standardizing of the spelling of names, localities, and taxonomic names.

Using General Refine Expression Language (GREL) and Jython, mass standardization is made easier and steps can be repeated across datasets.

Expression: `value.replace('M.N.B.','M.N. Bramletti').replace('W.P.W.','W.P. Woodring')`

row	value	value.replace('M.N.B.','M.N. B ...
27.	W.P.W.	W.P. Woodring
116.	Kew	Kew
168.	W.P.W., M.N.B.	W.P. Woodring, M.N. Bramlette
176.	M.N.B., W.P.W.	M.N. Bramlette, W.P. Woodring
212.	M.N.B.	M.N. Bramlette
766.	W.P. Irwin	W.P. Irwin

Using the GREL expression, only certain names were changed in the dataset.

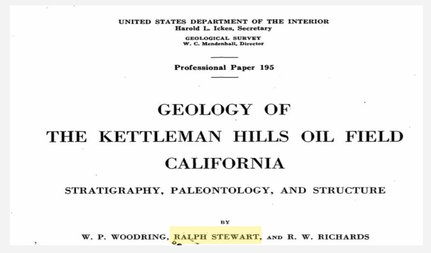


OpenRefine will filter all of the data from a single column and analyse patterns in the data to suggest changes.

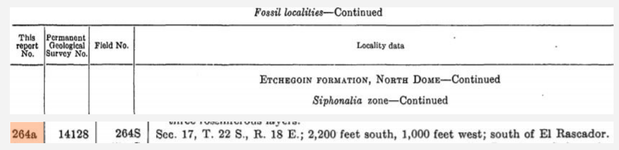
Enhancement: Data transcribed from labels can be lacking in context or information. OpenRefine provides tools for segmenting related data (i.e., collector name, taxonomic name, locality number) or finding recurring patterns. We can then use this information to reference outside resources such as ledgers, field books, publications, and records from other institutions for further data.

6 matching rows (1381 total)

	LocSiteNumber	LocSiteStation	Collector	Geography	Location Description	Taxo
1299.	USGS CENO LOC	14128	Stewart	Kings County, California	North Dome, Kettleman Hills, California, ! Zone, S. of El Rascal	Calyptrae Gabb
1300.	USGS CENO LOC	14128	Stewart	Kings County, California	North Dome, Kettleman Hills, California, Siphonalia*	Nassa cf
1301.	USGS CENO LOC	14128	Stewart	Kings County, California	North Dome, Kettleman Hills, California, siphonalia* zone	Nassa cf moraniana
1302.	USGS CENO LOC	14128	Stewart	Kings County, California	North Dome, Kettleman Hills, California, Siphonalia* Z.	Cathistop privosa*
1303.	USGS CENO LOC	14128	Stewart	Kings County, California	264 a R264S | North Dome, Kettleman Hills, California, Siphonalia Z.	Nassa mi iniguus
1304.	USGS CENO LOC	14128	Stewart	Kings County, California	Siphonalia Z. S. of Rascadon, 30-40 below BC & Mylex*	



A collector last name can lead to a publication and fullname. An unclear number can be found in a locality ledger.



This process enables more information about the locality and stratigraphy can be included with the transcribed label data to form a more complete occurrence record. It also provides context for information that would otherwise lack meaning.

An essential step in this workflow includes documentation of enhancements and interpretations within the record. We use notes and verbatim fields to maintain transparency and long-term understanding of the data.

5. Backfilling with Python

As data was processed we maintained separate master files for all persons and geologic age data in order to track variation and expand the data based on existing records.

After OpenRefine, each dataset is processed with a second Python script that fills in higher level geologic age information, fields documenting updates to this information, and links to existing records for all persons.

```
FP_GEOAGE = 'geologic_age2.csv'
FP_PARTIES = 'parties.csv'
FP_INPUT = 'NMNH-PALEO-20180126_AC_openrefine.csv'
FP_OUTPUT = os.path.basename(FP_INPUT).rsplit('_', 1)[0] + '_emuprep.csv'

def standardize(val):
    """Standardizes the format of the string to improve matching"""
    return val.strip().replace('.', '').replace(' ', '').lower()

def match_ages(records, errors=None):
    """Matches new data to geologic ages already in EMu"""
    if errors is None:
        errors = []
    # Read geologic age info from geologic_age2.csv
    ages = {}
    with open(FP_GEOAGE, 'r', encoding='utf-8') as gainput:
        rows = csv.reader(gainput, dialect='excel')
        keys = next(rows)
```

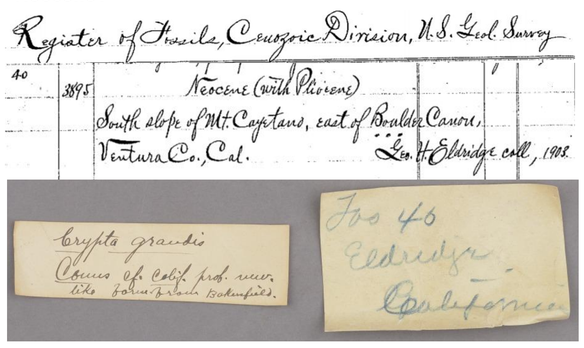
6. Ingest into EMu

Once the data is fully processed it is ingested into EMu, expanding the original skeleton records into more robust and standardized occurrence records with thorough documentation.

Prefi...	Catalog...	Event/Site: (Object Data from time of Collecting Event)/Locality: (Collection Details)/Event/Site: (Object/Remains Data)	Collector: (Versi...	Comments: (Identification Details)	Note: (Note Details)	System: (Geologic...	Series: (Geol
PAL	637989	Woodring : : USGS CENO LOC : 12656 : USA : California : Kings County : : : : Kettleman Hills | Avenal Gap Quadrangle | N end of South ...	Woodring		EMu record was cr...	Neogene	Pliocene
PAL	637990	Woodring : : USGS CENO LOC : 12657 : USA : California : Kings County : : : : Kettleman Hills | Avenal Gap Quadrangle | N end of South ...	Woodring	ID transcribed from label and not verified. Verbatim taxonomy: Neverita reclusiana (Desh...	EMu record was cr...	Neogene	Pliocene
PAL	637991	Woodring : : USGS CENO LOC : 12657 : USA : California : Kings County : : : : Kettleman Hills | Avenal Gap Quadrangle | N end of South ...	Woodring	ID transcribed from label and not verified. Verbatim taxonomy: Zirfaea cf. gabbi Tryon	EMu record was cr...	Neogene	Pliocene
PAL	637993	Woodring : : USGS CENO LOC : 12664 : USA : California : Kings County : : : : Kettleman Hills | Los Varones Quadrangle | E flank of Middl...	Woodring	ID transcribed from label and not verified. Verbatim taxonomy: Balanus	EMu record was cr...	Neogene	Pliocene
PAL	637997	Woodring : : USGS CENO LOC : 12657 : USA : California : Kings County : : : : Kettleman Hills | Avenal Gap Quadrangle | N end of South ...	Woodring	ID transcribed from label and not verified. Verbatim taxonomy: Ostrea sp.	EMu record was cr...	Neogene	Pliocene
PAL	637998	Woodring : : USGS CENO LOC : 12654 : USA : California : Kings County : : : : Kettleman Hills | Avenal Gap Quadrangle | W flank of South...	Woodring	ID transcribed from label and not verified. Verbatim taxonomy: Ostrea sp.	EMu record was cr...	Neogene	Pliocene
PAL	637999	Woodring : : USGS CENO LOC : 12657 : USA : California : Kings County : : : : Kettleman Hills | Avenal Gap Quadrangle | N end of South ...	Woodring	ID transcribed from label and not verified. Verbatim taxonomy: Pododesmus cf. macrosch...	EMu record was cr...	Neogene	Pliocene
PAL	638000	Woodring : : USGS CENO LOC : 12654 : USA : California : Kings County : : : : Kettleman Hills | Avenal Gap Quadrangle | W flank of South...	Woodring	ID transcribed from label and not verified. Verbatim taxonomy: Pseudocardium?	EMu record was cr...	Neogene	Pliocene
PAL	638001	R. Stewart : : USGS CENO LOC : 12798 : USA : California : Kings County : : : : Kettleman Hills | Avenal Gap Quadrangle | Kettleman Hills,...	R. Stewart		EMu record was cr...	Neogene	Pliocene
PAL	638005	R. Stewart : : USGS LOC : 12834 : USA : California : Kings County : : : : Kettleman Hills | Avenal Gap Quadrangle | Kettleman Hills, Avenal...	R. Stewart	ID transcribed from label and not verified. Verbatim taxonomy: Pseudocardium densatum ...	EMu record was cr...	Neogene	Pliocene
PAL	638007	R. Stewart : : USGS LOC : 12834 : USA : California : Kings County : : : : Kettleman Hills | Avenal Gap Quadrangle | Kettleman Hills, Avenal...	R. Stewart	ID transcribed from label and not verified. Verbatim taxonomy: D. cf. coalingensis	EMu record was cr...	Neogene	Pliocene
PAL	638013	M. Birkhauser : : UNKNOWN : 543 : USA : California : : : : Curry Mt. Autocline; on branch of creek along road near center of W line of s...	M. Birkhauser		EMu record was cr...		

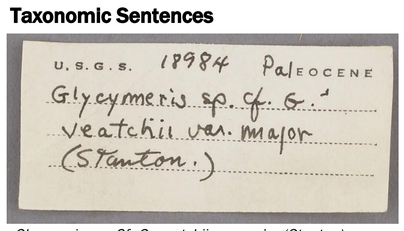
Fun with Data

Neocene?



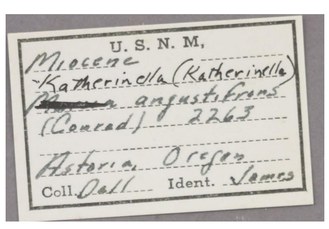
Effie Clark notations and housing

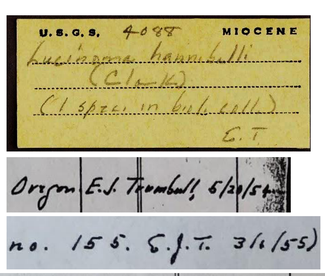
Taxonomic Sentences



Glycymeris sp. Cf. G. veatchii var. major (Stanton)

The many names of Ellen James Moore





Ficopsis sp cf. F remondii crascantensis Weaver

Conclusion and Next Steps

While processing this data it became even more apparent how important it is to clearly document what was done. When we see what was recorded over 100 years ago, but can't follow the work or understand a notation it means we essentially have to start from scratch. We don't want that for our successors.

This workflow enabled us to meet that challenge. We accounted for possibly unverified, poorly documented, or sparse analog data; utilized tools and procedures to efficiently transform this data into standardized vocabularies and structures; and we maintained transparency by clearly documenting the decisions and interpretations made during the process.

NEXT STEPS

We took an iterative approach to processing this data and have more work to do. A second pass to further parse information into more appropriate fields and to update outdated Taxonomy and Lithostratigraphy is in progress. This work includes:

- R package to compare outdated, messy taxonomy to datasets of currently accepted names
- Analysis of remaining unclear data across all related collections data for recognizable patterns
- More OpenRefine!!

EPICC Data Review Protocols

Last edit: 4/17/2018 by HL

Contents

Documentation of all protocols is important for consistent work across all datasets and staff and for ensuring long term understanding of how data was processed