CLAMS IN THE CITY AND SNAILS LOST AT SEA:
A FITNESS-FOR-USE ASSESSMENT OF AGGREGATED MARINE BIODIVERSITY DATA

AUSTIN HENDY

NATURAL HISTORY MUSEUM
LOS ANGELES COUNTY

Credit: **The Snail and the Whale**

PALEOCLIMATIC AND PALEOBIOGEOGRAPHIC IMPLICATIONS OF A PLEISTOCENE ASSEMBLAGE FROM ANGOLA, TROPICAL WEST AFRICA

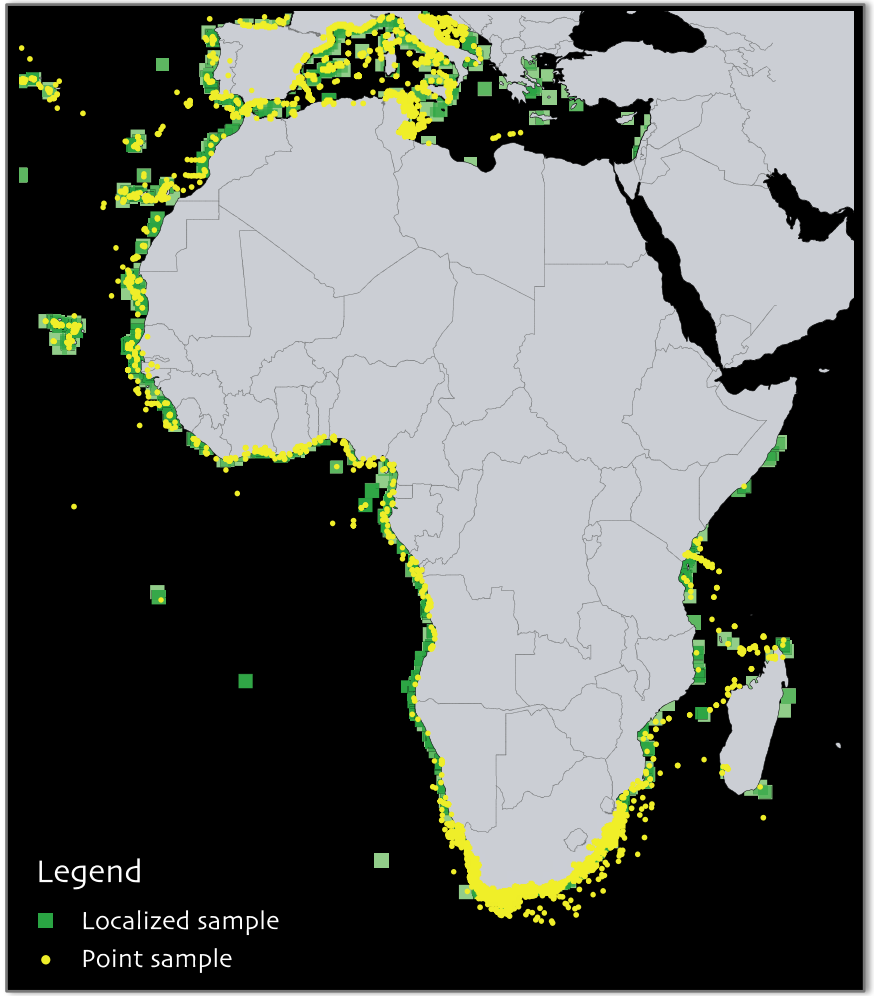
Jocelyn Sessa

Pedro Callapez

Pedro Dinis

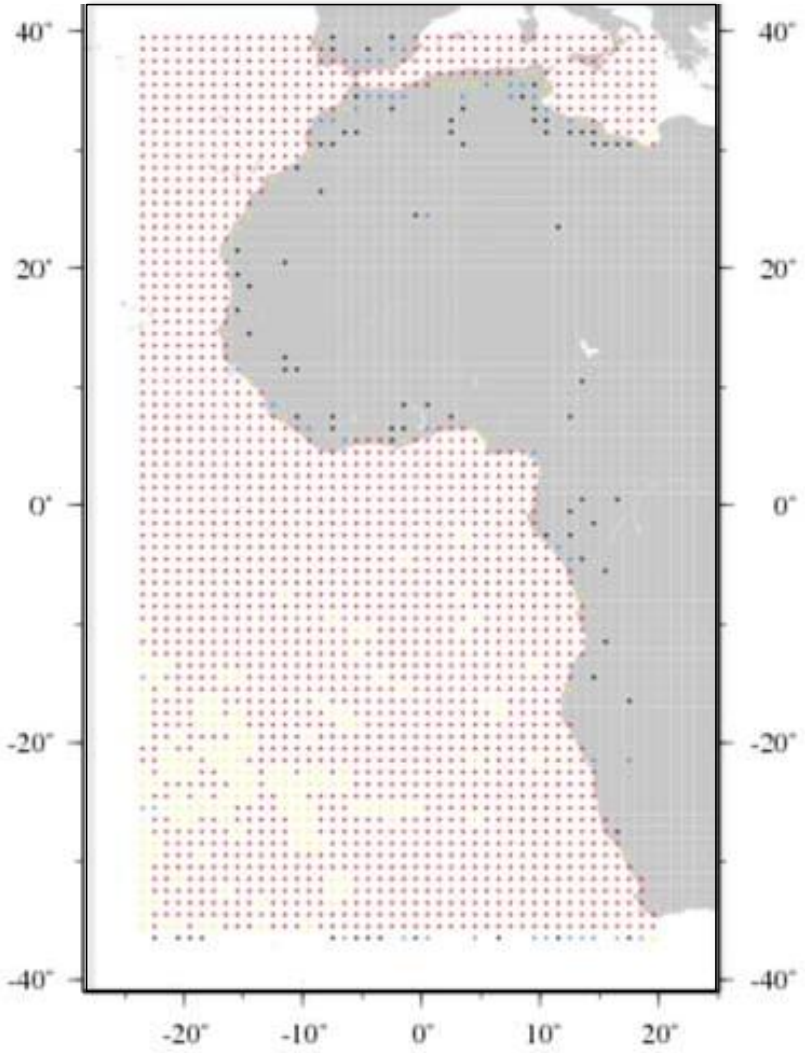
Austin Hendy

Occurrence data

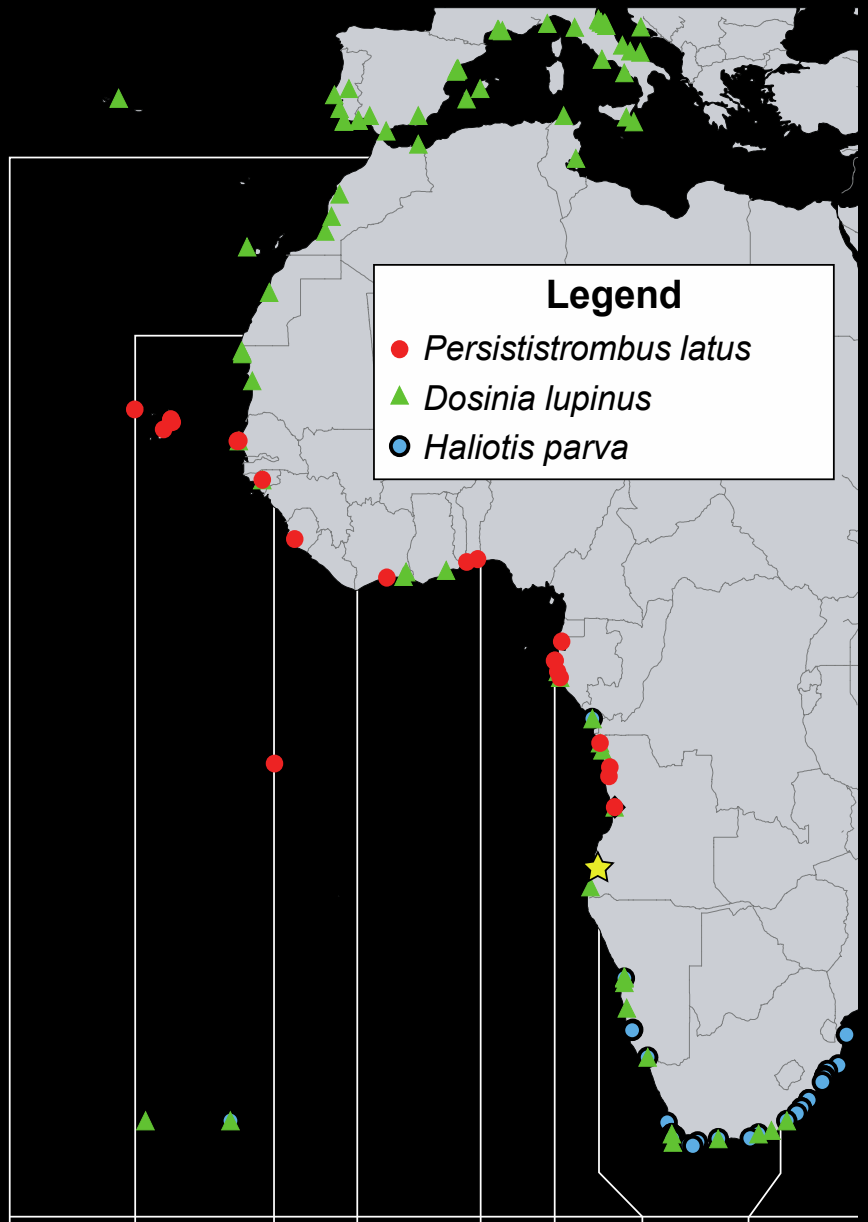


Occurrence data from published literature, museum & private collections, and data aggregators

Ocean data



Temperature/salinity data corresponding to these geographic coordinates were downloaded from the World Ocean Database



Warmest	26-28	27-29	28	26-27	28	27.5	18-20	25-27
Coldest	16-18	18-19	26	22-24	25	19.5	10-12	20-22


Google earth
Data SIO, NOAA, U.S. Navy, NGA, GEBCO
Image Landsat / Copernicus

DATA SOURCES

Data were downloaded from:

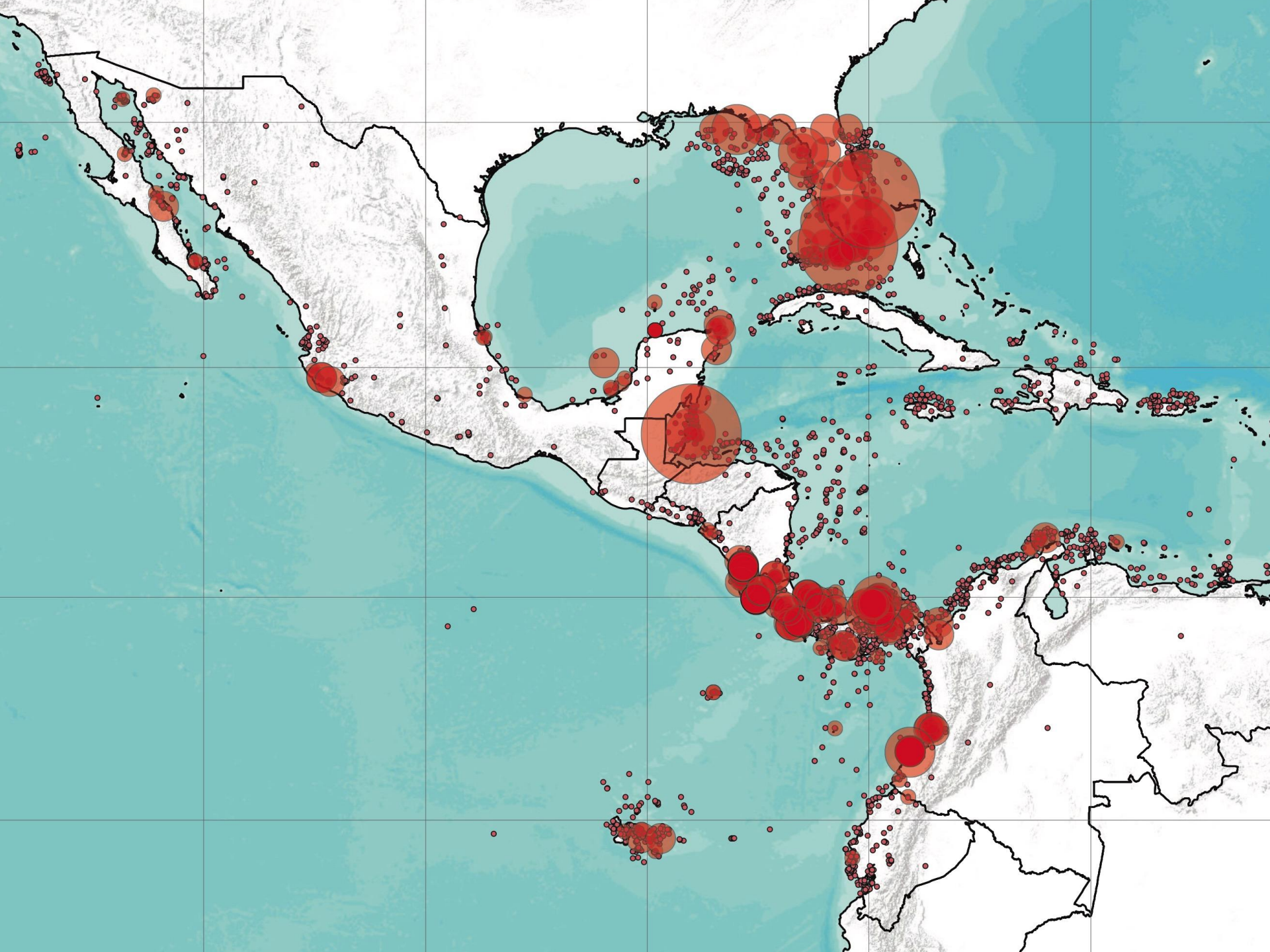
- GBIF (Global Biodiversiy Information Facility)
- iDigBio
- OBIS (Ocean Biogeographic Information System)

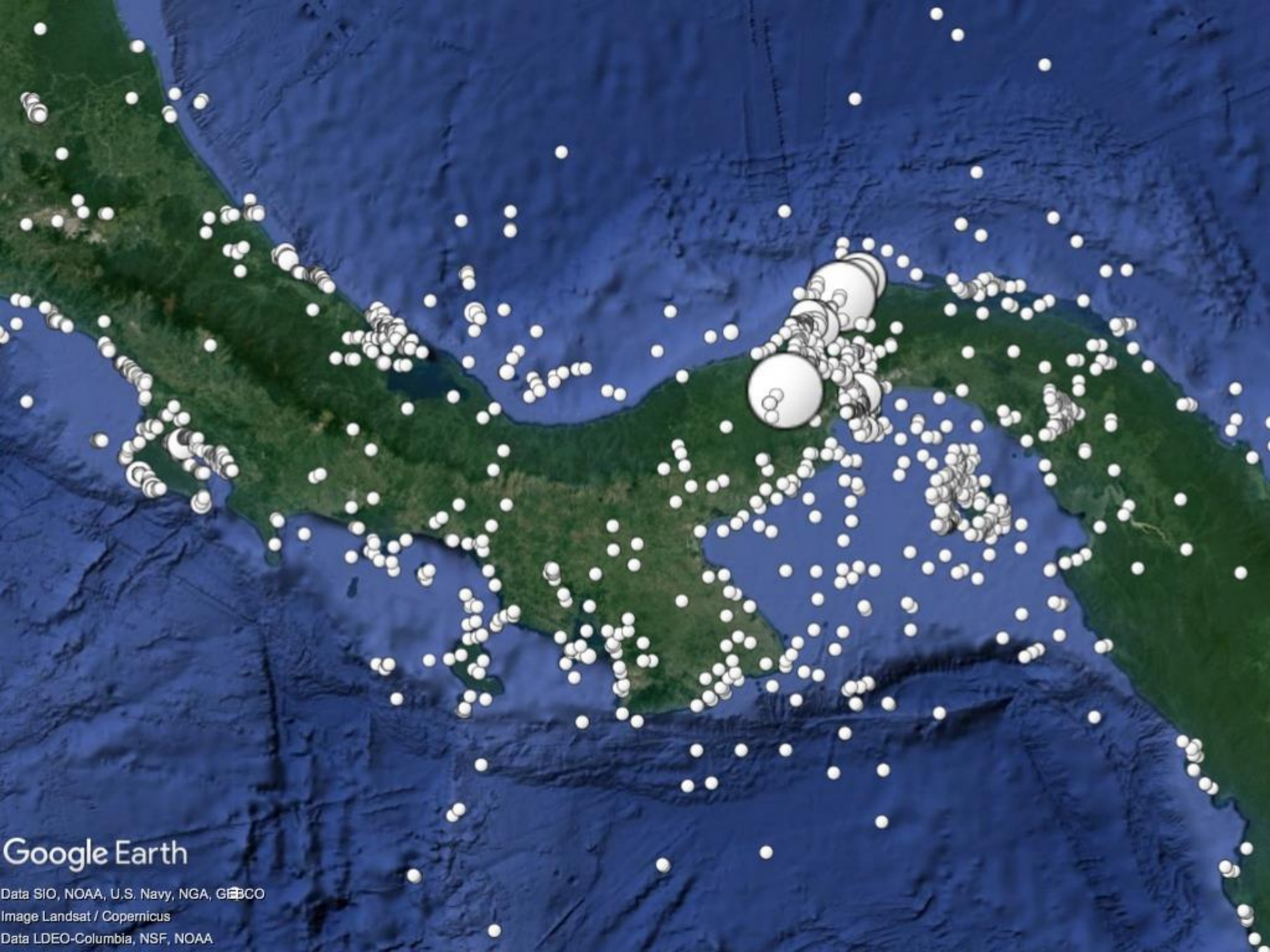
Georeferencing checked with:

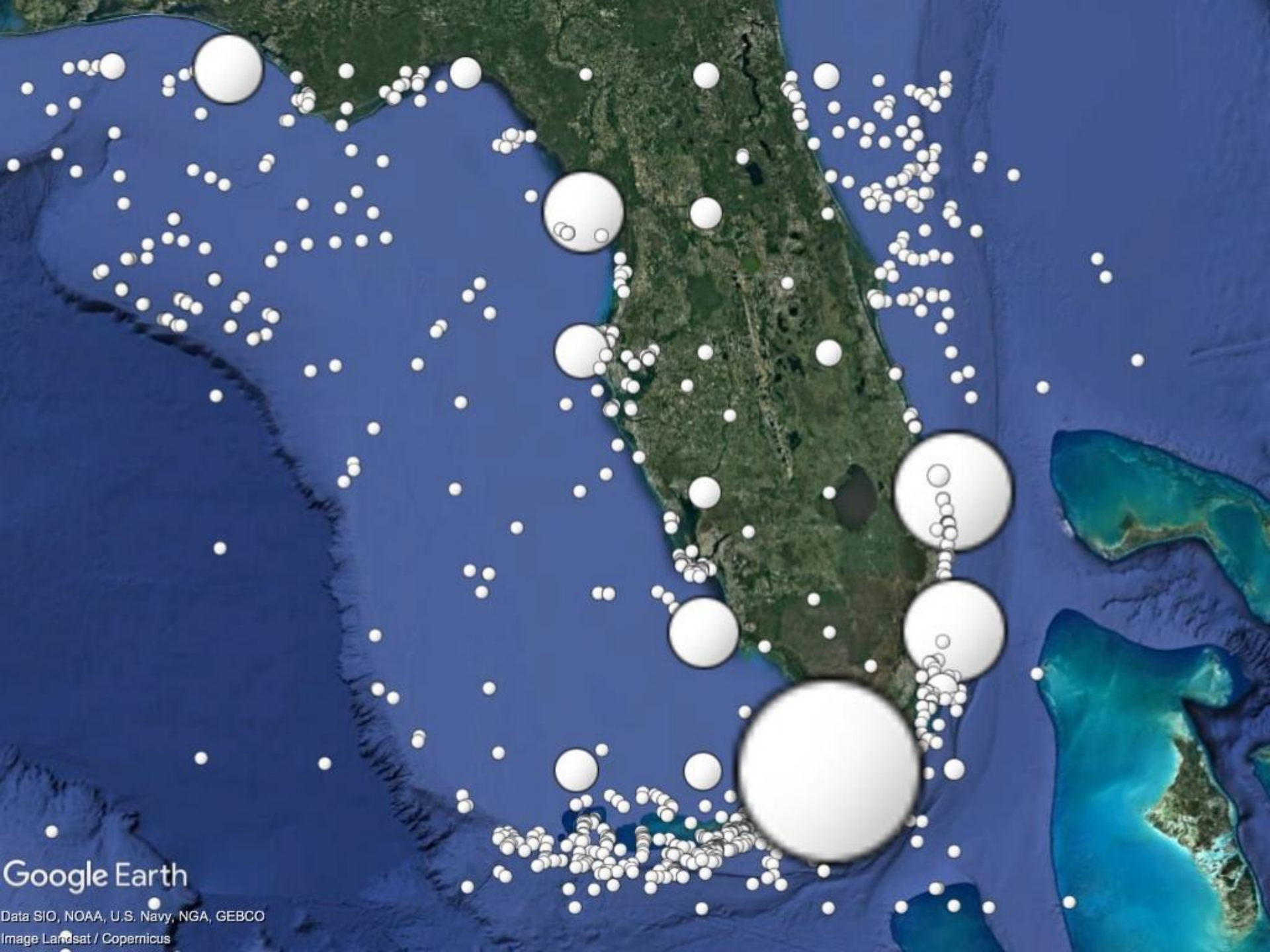
- Manual labor (Google Earth Pro, eyeballs, & common sense)

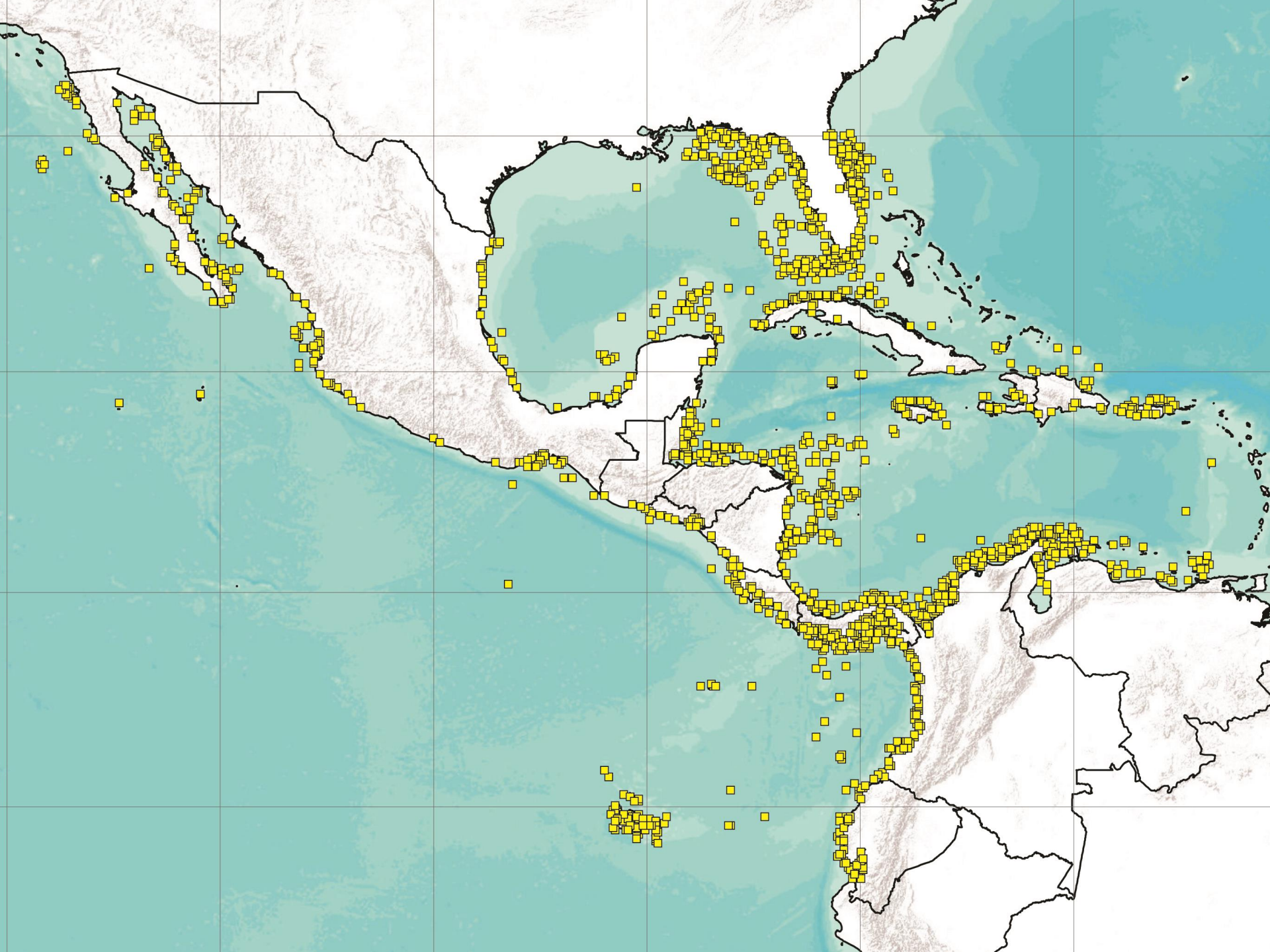
Data matches with:

- WoRMS (taxonomy)
- Marine Regions Gazetteer (marine georeferenced place names & biogeographic boundaries)
- Bio-ORACLE (marine data layers for ecological modelling)




Google Earth
Data SIO, NOAA, U.S. Navy, NGA, GEBCO
Image Landsat / Copernicus
Data LDEO-Columbia, NSF, NOAA


Google Earth
Data SIO, NOAA, U.S. Navy, NGA, GEBCO
Image Landsat / Copernicus



RESULTS

Of >126,000 occurrences screened for errors

- 61% were georeferenced (>77,000 occurrences)
- 38% of those occurrences not georeferenced **could be** accurately georeferenced
- 1% (1260 occurrences) **could not be** accurately georeferenced.

- **Of those georeferenced, 28% (>22,000 occurrences) were incorrect**

  Coordinates inconsistent with verbatim geographic descriptions, relative to coastlines, or country or ocean assignments

RESULTS

Three common types of gereferencing errors were observed:

1. Use of an inappropriate geographic/political **CENTROID** (e.g., centroid of a country, state, city)

   37% (8100 occurrences)

2. **ROUNDING** to nearest degree of latitude and longitude

   21% (4700 occurrences)

3. Simply **POOR ESTIMATION** of coordinates; *a catch-all for unfathomable georeferencing*

   40% (8800 occurrences)

RESULTS

Relatively few institutions are responsible for the majority of the errors

- Three U.S. institutions and one European museum had error rates in excess of 20% of their aggregated records
- The Field Museum had an error rate less than 1%

Types of errors are not evenly distributed across institutions

- One institution with 2685 of 3888 incorrectly georeferenced coordinates due to use of **inappopriate centroids**
- Another institution with 8738 of 14108 incorrectly georeferenced coordinates due to **rounding**

RESULTS

Error rates for georeferencing of occurrences from Latin American countries are much higher among U.S. & European insitutions

- Occurrences from **Costa Rica**
  - Local institutions ->0.7%
  - U.S.+European institutions ->18.3%
- Occurrences from **Mexico**
  - Local institutions -> 8.9%
  - U.S.+European institutions -> 47.3%
- Occurrences from **Ecuador**
  - U.S.+European institutions -> 23.6%

RESULTS

Local institutions also achieve higher rates of georeferencing than U.S. & European institions

- Occurrences from **Colombia**
  - Local institutions ->99.4%
  - U.S.+European institutions ->31.5%
- Occurrences from **Costa Rica**
  - Local institutions -> 99.9%
  - U.S.+European institutions -> 38.4%
- Occurrences from **Mexico**
  - Local institutions -> 100%
  - U.S.+European institutions -> 76.6%

Melongena corona

Correct: 26
New: 2
Incorrect: 87

Centroid: 9
Rounded: 72
Poorly Estimated: 6
Transposed: 0
Average error distance: 54.9 km

Biogeographic zone: 70 -> 99
Temperature range: 15.4-31.7° -> 15.2-31.9

Conus jaspideus

Correct: 27
New: 52
Incorrect: 82

Centroid: 23
Rounded: 40
Poorly Estimated: 14
Transposed: 5
Average error distance: 826 km (66.3 km)

Biogeographic zone: 82 -> 147
Temperature range: 15.4-32.6° -> 15.4-33.6

Strombus alatus

Correct: 17
New: 4
Incorrect: 51

Centroid: 8
Rounded: 38
Poorly Estimated: 5
Transposed: 0
Average error distance: 72.5 km

Biogeographic zone: 49 -> 54
Temperature range: 15.6-32.2° -> 15.4-31.8

Conus purpurascens

Correct: 74
New: 192
Incorrect: 64

Centroid: 36
Rounded: 2
Poorly Estimated: 26
Transposed: 0
Average error distance: 55.9 km

Biogeographic zone: 85 -> 268
Temperature range: 17.4-33.6° -> 17.5-32.5

EXPLANATIONS

Poor estimation and use of inappropriate centroids

- Occur primarily in some of the earliest institutions to make their data available online
- The tools and training were not available at this time.
- Lack of realization of how these data would be used in the future

Coordinate rounding

- Primarily occurs at one institution
- Likely a collection policy (to cloak data) or a quick and easy way to georeference large numbers of localities

Accuracy of local institutions

- Collections composed from field sampling rather than donations
- Knowledge of local geography/oceanography and language

CONSQUENCES

Errors in aggregated data erodes community confidence in all available data!

Three areas require attention:

- Improvement of revision and republication methods for data publishers;
- New and improved methods for documenting different areas of geospatial fitness-for-use;
- Adoption of new technology to increase the speed at which fitness-for-use enhancement can be performed on available data.

GBIF. 2010. GBIF Position Paper on Future Directions and Recommendations for Enhancing Fitness-for- Use Across the GBIF Network, version 1.0. authored by Hill, A. W., Otegui, J., Ariño, A. H., and R. P. Guralnick. 2010

RECOMMENDATIONS

- Use of DWC fields for georeferencing so that downstream users are aware methods, sources, and uncertainties
  - Poor use of protocol (20%), data sources (4%), verification status (4%), and uncertainty (8%) ***data providers***
- Revision and republication of data by relatively few institutions
  - Two institutions responsible for 75% of the incorrectly georeferenced localities ***data providers***
- Improved methods for documenting geospatial data quality
  - Flagging records for researachers and providing feedback to data providers ***data aggregators***
- Greater awareness and development of standards, tools and worfklows for georeferencing marine collecting events.

data aggregators & funding agencies

THANK YOU

Credit: **The Onion**

Credit: **The Snail and the Whale**