Georeferencing Fish Occurrences

A Community-Based Model

Henry L. Bart Jr.
&
Nelson E. Rios
Tulane University Biodiversity Research Institute
Challenge

• Assigning geographic coordinates to hundreds of thousands of newly digitized collection records from across the globe produced by different TCNs.
• Many samples of even different kinds of organisms taken from the same locality.
• How can projects avoid duplicating effort georeferencing these?
• How can georeferencing process be efficiently managed, regionally specialized, broadened in participation?
Solution

- GEOLocate’s Collaborative Georeferencing (CoGe) Platform.
- Presenting results of 2012 CSBR project involving fish collection data.
- First large scale test of CoGe system.
- Preliminary results of experiments in crowd-sourced georeferencing.
FishNet 2

global network

71 data providers

3.7 million lots
35+ million specimens
57% georeferenced
Objectives

Expand the number of records within FishNet 2 to approximately 4 million lots representing over 30 million specimens

Georeference all records lacking geographic coordinates

Repatriate results to source data providers

321,102 collecting events, representing 1,568,695 specimen lots, in need of georeferencing in FishNet2
Georeferencing FishNet 2

- 250,000 collecting events
- 18 months
- 12 technicians
- 12+ volunteers
- GEOLocate’s CoGe System
- Crowdsourcing experiments
The Dirty Dozen
Workflow

- Harvest data by Institution from FishNet 2
- Standardize higher geographies
- Partition into regional datasets
- Upload to data portal
- Assign datasets to technicians
- GEO Locate
- Review results
- Repatriate
Georeferencing Status

11 January – 14 October 2013
727K Specimens / 168K Localities / 9 Months (67%)
Community-Wide
Daily Rates of Localities Georeferenced (excluding volunteer effort)

11 January – 14 October 2013
Cumulative Results by Institution

Efficiency of collaborative approach: regional specialization

11 January – 14 October 2013
Efficiencies of the Collaborative System: Simultaneous Correction of Multiple Georeferences

High percentage (19.6%) of “verification events” involved multiple georeferences corrected at same time.
Georeferencing Output

Base Map - 1:2M 8-Digit HUCs
Current tools targeted at “experts”
How Do Non-Experts Perform?
Crowdsourcing Experiment

- 7,463 lots (916 localities) with coordinates already determined randomly selected from the state of Arkansas.
  - Data provided by KU, MMNS, NCSM, OKMNH, OSUM, ROM, TU, USNM, YPM via FishNet

- Group into 5 regional datasets
- Each region assigned to a Tulane student for georeferencing
- Evaluated accuracy of student results
Georeferencing Profiles

Student 1
N=278 localities

Student 2
N=223 localities

Student 3
N=163 localities

Student 4
N=116 localities

Student 5
N=136 localities
Expert vs. Volunteer Comparison

**Experts:**
- 163,586 locality verifications
  - 143,108 corrected
  - 20,478 skipped (12.5% rate)
- 703,794 specimen records
  - 635,251 corrected
  - 68,543 skipped (9.7% rate)
- 10 localities per hour; high accuracy

**Volunteers:**
- 4,558 locality verifications
  - 4,125 corrected
  - 433 skipped (9.5% rate)
- 23,026 specimen records
  - 21,541 corrected
  - 1,485 skipped (6.9% rate)
- 22 localities per hour; low accuracy
Contributors to Volunteer Georeferencing Inaccuracy

- **Inexperience** in Georeferencing
- **Lack of geographic familiarity** with regions
- No access to **supporting data** (catalogs, field notes, ship logs, etc.)
- Specificity assessments
  - 6 miles west of Dierks
- **Errors in original data**
  - Original data still need to be double checked (likely reason for large inaccuracies)
- Student **motivation varies** over time
Crowdsourcing: Conceptual Requirements

• Data Quality
  – Multiple verifications per locality
  – Periodic skills assessments using known controls
  – Reputation-based scoring
  – Validation against higher geography & taxon distributions
  – Spatial clustering & outlier removal
  – Batch Review & refinement by experts or promoted citizen scientists

• Engaging K-16, General Public
  – Allow georeferencers to select projects or regions
  – User promotion and reward system
  – Client software must present a low barrier of entry
  – Rich educational experience beyond moving points on a map
Adding Crowdsourcing Components to GEOLocate’s Collaborative Georeferencing Architecture (pending Biofinder proposal).
## Acknowledgements

### Institutional Collaborators

<table>
<thead>
<tr>
<th>Andy Bentley</th>
<th>Luiz Rocha</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dave Catania</td>
<td>Rob Robins</td>
</tr>
<tr>
<td>Michael Doosey</td>
<td>Mark Sabaj</td>
</tr>
<tr>
<td>Rick Feeney</td>
<td>Leo Smith</td>
</tr>
<tr>
<td>Bill Fink</td>
<td>Alexandra Snyder</td>
</tr>
<tr>
<td>John Friel</td>
<td>Chris Taylor</td>
</tr>
<tr>
<td>Dean Hendrickson</td>
<td>Christine Thacker</td>
</tr>
<tr>
<td>Amy McCune</td>
<td>Tom Turner</td>
</tr>
<tr>
<td>John Lundberg</td>
<td>Mark Westneat</td>
</tr>
<tr>
<td>Douglas Nelson</td>
<td>Ed Wiley</td>
</tr>
<tr>
<td>Larry Page</td>
<td></td>
</tr>
</tbody>
</table>

### Georeferencing Technicians

<table>
<thead>
<tr>
<th>Diego Barroso</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Melissa Casarez</td>
<td></td>
</tr>
<tr>
<td>Paul DeSalles</td>
<td></td>
</tr>
<tr>
<td>Justin Grubich</td>
<td></td>
</tr>
<tr>
<td>Mariangeles Arce Hernandez</td>
<td></td>
</tr>
<tr>
<td>Estella Hernandez</td>
<td></td>
</tr>
<tr>
<td>Benito Lorenzo</td>
<td></td>
</tr>
<tr>
<td>Theresa Lorraine Mclnnes</td>
<td></td>
</tr>
<tr>
<td>Laura Porturas</td>
<td></td>
</tr>
<tr>
<td>Megan Roberson</td>
<td></td>
</tr>
<tr>
<td>Michelle Vanderwel</td>
<td></td>
</tr>
<tr>
<td>Rachel Vinsel</td>
<td></td>
</tr>
<tr>
<td>Katy Wichman</td>
<td></td>
</tr>
<tr>
<td>INHS &amp; TU Student Volunteers</td>
<td></td>
</tr>
</tbody>
</table>

This material is based in part upon work supported by the National Science Foundation under Grant Numbers DBI-0852141 and DBI-1202953. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.