The BELS Georeference Matcher

TaxonWorks Together
26 Oct 2023
The **BELS Georeference Matcher**

**Biodiversity Enhanced Location Services**

TaxonWorks Together

26 Oct 2023

Julie Allen
Michael Denslow
Ed Gilbert
Rob Guralnick
Rafe LeFrance
Nelson Rios
John Wieczorek
Paula Zermoglio
Back in May 2020…

Imagining a Global Gazetteer of Georeferences

Paula Zermoglio - VertNet
Rob Guralnick - University of Florida
Julie Allen - University of Nevada Reno
Has someone else already georeferenced this location well enough that I can use it?
Has someone else already georeferenced this location well enough that I can use it?
Has someone else already georeferenced this location well enough that I can use it?
Darwin Core terms

Classes

Simple Darwin Core
- Record & Dataset
- Occurrence
- Organism
- Material Sample
- Event
- Location
- Geological Context
- Identification
- Taxon

Auxiliary classes
- Resource Relationship
- Measurement or Fact
- Chronometric Age
<table>
<thead>
<tr>
<th>Category</th>
<th>Key Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identifiers</td>
<td>locationID, higherGeographyID</td>
</tr>
<tr>
<td>Geographic Data</td>
<td>higherGeography, islandGroup, island, waterBody, continent, country, countryCode, stateProvince, county, municipality, locality, verbatimLocality, minimumElevationInMeters, maximumElevationInMeters, minimumDepthInMeters, maximumDepthInMeters, minimumDistanceAboveSurfaceInMeters, maximumDistanceAboveSurfaceInMeters, verbatimElevation, verbatimDepth, decimalLatitude, decimalLongitude, coordinatePrecision, verbatimLatitude, verbatimLongitude, verbatimCoordinates, coordinateUncertaintyInMeters, geodeticDatum</td>
</tr>
<tr>
<td>Vertical Components</td>
<td></td>
</tr>
<tr>
<td>Georeference</td>
<td>footprintSRS, footprintSpatialFit, pointRadiusSpatialFit, verbatimCoordinateSystem, footprintWKT, verbatimSRS</td>
</tr>
<tr>
<td>Georeference Metadata</td>
<td>georeferencedBy, georeferenceProtocol, georeferencedDate, georeferenceSources, georeferenceRemarks</td>
</tr>
<tr>
<td>Other Data</td>
<td>locationAccordingTo, locationRemarks</td>
</tr>
</tbody>
</table>
Locations

- Many distinct location descriptions (strings) refer to the same place.
Locations

- Many distinct location descriptions (strings) refer to the same place.
- Usually multiple specimens and/or observations (even of different taxa) have the same location descriptions.
• Many distinct location descriptions (strings) refer to the same place.

• Usually multiple specimens and/or observations (even of different taxa) have the same location descriptions.

• Sometimes one or more among the many location descriptions of a place have a georeference.
Has someone else already georeferenced this location well enough that I can use it?
Location terms

identifiers

geographic data

vertical components

georeference

georeference metadata

other data

decimalLatitude, decimalLongitude, coordinatePrecision
verbatimLatitude, verbatimLongitude, verbatimCoordinates
coordinateUncertaintyInMeters, geodeticDatum
footprintSRS, footprintSpatialFit, pointRadiusSpatialFit
verbatimCoordinateSystem, footprintWKT, verbatimSRS
georeferencedBy, georeferenceProtocol, georeferencedDate
georeferenceSources, georeferenceRemarks
locationAccordingTo, locationRemarks
<table>
<thead>
<tr>
<th>Location terms</th>
<th>identifiers</th>
<th>geographic data</th>
<th>vertical components</th>
<th>georeference metadata</th>
<th>other data</th>
</tr>
</thead>
<tbody>
<tr>
<td>locationID</td>
<td>higherGeographyID</td>
<td>geographic data</td>
<td>vertical components</td>
<td>georeference metadata</td>
<td>other data</td>
</tr>
<tr>
<td>higherGeography</td>
<td>islandGroup</td>
<td>island</td>
<td>waterBody</td>
<td></td>
<td></td>
</tr>
<tr>
<td>continent</td>
<td>country</td>
<td>countryCode</td>
<td>stateProvince</td>
<td></td>
<td></td>
</tr>
<tr>
<td>county</td>
<td>municipality</td>
<td>locality</td>
<td>verbatimLocality</td>
<td></td>
<td></td>
</tr>
<tr>
<td>minimumElevationInMeters</td>
<td>maximumElevationInMeters</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>minimumDepthInMeters</td>
<td>maximumDepthInMeters</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>minimumDistanceAboveSurfaceInMeters</td>
<td>verbatimElevation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>maximumDistanceAboveSurfaceInMeters</td>
<td>verbatimDepth</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>decimalLatitude</td>
<td>decimalLongitude</td>
<td>coordinatePrecision</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>verbatimLatitude</td>
<td>verbatimLongitude</td>
<td>verbatimCoordinates</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coordinateUncertaintyInMeters</td>
<td>geodeticDatum</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>footprintSRS</td>
<td>footprintSpatialFit</td>
<td>pointRadiusSpatialFit</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>verbatimCoordinateSystem</td>
<td>footprintWKT</td>
<td>verbatimSRS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>georeferencedBy</td>
<td>georeferenceProtocol</td>
<td>georeferencedDate</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>georeferenceSources</td>
<td>georeferenceRemarks</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>locationAccordingTo</td>
<td>locationRemarks</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Has someone else already georeferenced this location well enough that I can use it?
## Fitness for Use

<table>
<thead>
<tr>
<th>Usefulness</th>
<th>Occurrences</th>
<th>%</th>
<th>Locations</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>-</td>
<td>2,232,326,955</td>
<td>174,245,784</td>
<td></td>
</tr>
<tr>
<td>Coordinates</td>
<td>mappable</td>
<td>2,093,146,781</td>
<td>149,565,869</td>
<td></td>
</tr>
</tbody>
</table>

Data from GBIF snapshot 2022-07-14. Distinct Locations considering all terms in the Darwin Core Location class.
## Fitness for Use

<table>
<thead>
<tr>
<th>Usefulness</th>
<th>Occurrences</th>
<th>%</th>
<th>Locations</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total</strong></td>
<td>-</td>
<td>2,232,326,955</td>
<td>174,245,784</td>
<td></td>
</tr>
<tr>
<td><strong>Coordinates</strong></td>
<td>mappable</td>
<td>2,093,146,781</td>
<td>149,565,869</td>
<td>85.8</td>
</tr>
<tr>
<td><strong>Coordinates + uncertainty</strong></td>
<td>mappable with circle</td>
<td>717,870,489</td>
<td>91,491,38</td>
<td>48</td>
</tr>
</tbody>
</table>

Data from GBIF snapshot 2022-07-14. Distinct Locations considering all terms in the Darwin Core Location class.
## Fitness for Use

<table>
<thead>
<tr>
<th></th>
<th>Usefulness</th>
<th>Occurrences</th>
<th>%</th>
<th>Locations</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total</strong></td>
<td>-</td>
<td>2,232,326,955</td>
<td>174,245,784</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Coordinates</strong></td>
<td>mappable</td>
<td>2,093,146,781</td>
<td>93.8</td>
<td>149,565,869</td>
<td>85.8</td>
</tr>
<tr>
<td><strong>Coordinates + uncertainty</strong></td>
<td>mappable with circle</td>
<td>717,870,489</td>
<td>32.2</td>
<td>91,491,384</td>
<td>48</td>
</tr>
<tr>
<td><strong>Coordinates + uncertainty + identifiable datum</strong></td>
<td>minimally complete</td>
<td>597,054,795</td>
<td>26.7</td>
<td>81,320,951</td>
<td>46.7</td>
</tr>
</tbody>
</table>

Data from GBIF snapshot 2022-07-14. Distinct Locations considering all terms in the Darwin Core Location class

Fitness for Use

Distinct Locations

N=174,245,784

53.3% with georeference
46.7% without georeference

Data from GBIF snapshot 2022-07-14. Distinct Locations considering all terms in the Darwin Core Location class
## Fitness for Use

<table>
<thead>
<tr>
<th>Usefulness</th>
<th>Occurrences</th>
<th>%</th>
<th>Locations</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total</strong></td>
<td>-</td>
<td>2,232,326,955</td>
<td>174,245,784</td>
<td></td>
</tr>
<tr>
<td>Coordinates</td>
<td>mappable</td>
<td>2,093,146,781</td>
<td>93.8</td>
<td>149,565,869</td>
</tr>
<tr>
<td>Coordinates + uncertainty</td>
<td>mappable with circle</td>
<td>717,870,489</td>
<td>32.2</td>
<td>91,491,380</td>
</tr>
<tr>
<td>Coordinates + uncertainty + identifiable datum</td>
<td>minimally complete</td>
<td>597,054,795</td>
<td>26.7</td>
<td>81,320,951</td>
</tr>
<tr>
<td>Coordinates + Uncertainty + identifiable datum + source + protocol</td>
<td>theoretically reproducible</td>
<td>9,095,539</td>
<td>0.41</td>
<td>1,465,243</td>
</tr>
</tbody>
</table>

Data from GBIF snapshot 2022-07-14. Distinct Locations considering all terms in the Darwin Core Location class.
Web Application

Biodiversity Enhanced Location Services (BELS) - Georeference Matcher

Upload a comma-separated input file that contains location information. Choose an email address to which to send the notification when the results are ready. Choose an output file name. This name will form an identifying part of the results file name, which will be a gzipped CSV file or files with an extension .csv.gz added.

Choose File  No file chosen

Notification email address

Output file name only (output will be gzipped CSV)

Submit

https://localityservice.uc.r.appspot.com/
How does it work?

- Gazetteer of shared Locations
  - Process for matching strings
  - Compute best georeference
How does it work?

- Gazetteer of shared Locations
  - Process for matching strings
  - Compute best georeference

- Georeference Matcher
  - Process for matching strings
  - Find best georeference
Gazetteer

- Import Occurrences into Google BigQuery
- Assign unique identifier based on Location term contents
- Standardize countryCode
- Select for valid coordinates
- Select for valid coordinate uncertainty
- Standardize coordinate precision
- Interpret geodeticDatum
- Calculate georeference score
Matching str 3 includes: higher geography (sans continent country, with interpreted countryCode), locality (collapse with verbatimLocality), elevation, and depth.

Matching str 2 includes str 3 plus verbatim coordinate terms.

Matching str 1 includes str 2 plus decimal coordinate terms.
- removeSymbols(): Remove punctuation and symbols except . , / - and +
- saveNumbers(): Replace , . / - and + with space except between digits
- simplifyDiacritics(): Normalize unicode, remove white space, lowercase, and change diacritics to ASCII "equivalents"
**Gazetteer**

**A**

All Distinct Locations

- GBIF.org
- iDigBio
- VernNet

**B**

<table>
<thead>
<tr>
<th>GAZETTEER of GEOREFERENCEs</th>
</tr>
</thead>
<tbody>
<tr>
<td>hg + locality + other</td>
</tr>
<tr>
<td>verbatim coords</td>
</tr>
<tr>
<td>decimal coords</td>
</tr>
<tr>
<td>Datum</td>
</tr>
<tr>
<td>Unc</td>
</tr>
<tr>
<td>georef metadata</td>
</tr>
</tbody>
</table>

**C**

Normalization

- str 1
- str 2
- str 3

**D**

Choosing best georeferences

- Best str 1
- Best str 2
- Best str 3

If no result:

- Best Result
- Best Result
- Best Result

**BEST RESULT**
Find the best georeference

A
Max uncert = a → keep candidate A
Bounding box N-S > 2a
=> DISCARD

B
Max uncert = a = b → keep candidates A, B
Bounding box N-S, E-W ≤ 2a, 2b → keep candidates A, B

\[ \overline{AX} < \overline{BX} \] → keep candidate A
\[ a \geq AB, AC \]
=> KEEP A

C
Max uncert = a → keep candidate A
Bounding box N-S, E-W ≤ 2a → keep candidate A

\[ \overline{AX} = \min \] → keep candidate A
\[ a < AC \]
=> DISCARD

D
Max uncert = a = b → keep candidates A, B
Bounding box N-S, E-W ≤ 2a, 2b → keep candidates A, B

\[ \overline{AX} = \overline{BX} \] → keep candidates A, B
\[ a \geq AB, b \geq BA \] → keep candidates A, B
=> KEEP result with best metadata
For any given string, in order:

1. Its uncertainty must be equal to the maximum uncertainty in the set of possible georeferences.

2. The distance of its center to the centroid of all the georeference centers in the set must be equal to the minimum distance to the centroid among all the candidates from a) (i.e., the center has to be closest or tied for closest to the centroid of all the georeferences that have the maximum uncertainty).

3. The distance from its center to any other georeference center in the set must not exceed the maximum uncertainty (i.e., the candidate must contain the centers of all the other georeferences in the set).

4. If multiple choices still remain after the preceding criteria, prioritize by the pre-established criteria for best georeference metadata.

5. If multiple choices still remain, each is as good as any other, so we select the first georeference in the list.
Georeference Matcher Output

**Best Result fields:**
- bels_match_country
- bels_interpreted_countrycode
- bels_matchwithcoords
- bels_matchverbatimcoords
- bels_matchesanscoords
- bels_decimallatitude
- bels_decimallongitude
- bels_geodeticdatum
- bels_coordinateuncertaintyinmeters
- bels_georeferencedby
- bels_georeferenceddate
- bels_georeferenceprotocol
- bels_georeferencesources
- bels_georeferenceremarks
- bels_georeference_score
- bels_georeference_source
- bels_best_of_n_georeferences
- bels_match_type
Proof of concept

- Nitrogen-fixing plants
  - record source: GBIF
  - ~ 40,000 species
  - > 33M occurrences
Proof of concept

- Matching:
  - best practice georeferences
  - exact match only
  - no extra tricks
Proof of concept

# occurrences N-fixing species

Before

Data from GBIF snapshot 2019-04-15.

\[ N \approx 10.2M \]

31%

69%

\[ N \approx 21.4M \]

green with georeference

blue without georeference
Proof of concept

Data from GBIF snapshot 2019-04-15.

# occurrences N-fixing species

- N ≈ 18.0M
- N ≈ 3.6M
- N ≈ 10.2M

58%
31%
11%

Data from GBIF snapshot 2019-04-15.

With georeference
Without georeference, but retrievable
Without georeference and not retrievable
API Integration

Symbiota

GEO Locate

Georeference Matcher
The BELS Georeference Matcher

Thank you!

Digi-Leap

Julie Allen
Michael Denslow
Ed Gilbert
Rob Guralnick
Rafe LeFrance
Nelson Rios
John Wieczorek
Paula Zermoglio