The BELS Georeference Matcher

TaxonWorks Together
26 Oct 2023

Digi-Leap

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

The BELS Georeference Matcher

Biodiversity Enhanced Location Services

TaxonWorks Together
26 Oct 2023

Digi-Leap

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

Question

Has someone else already georeferenced this location well enough that I can use it?

Question

Has someone else already georeferenced this location well enough that I can use it?

Sources

GBIF.org

iDigBio
Integrated Digitized Biocollections

VertNet

Question

Has someone else already georeferenced this location well enough that I can use it?


Liparia splendens splendens
Photo by flamelily, CC0

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

Location terms

locationID
higherGeographyID
higherGeography
islandGroup
island
waterBody
continent
country
countryCode
stateProvince
county
municipality
locality
verbatimLocality
minimumElevationInMeters
maximumElevationInMeters
minimumDepthInMeters
maximumDepthInMeters
minimumDistanceAboveSurfaceInMeters
verbatimElevation
maximumDistanceAboveSurfaceInMeters
verbatimDepth
decimalLatitude
decimalLongitude
coordinatePrecision
verbatimLatitude
verbatimLongitude
verbatimCoordinates
coordinateUncertaintyInMeters
geodeticDatum
footprintSRS
footprintSpatialFit
pointRadiusSpatialFit
verbatimCoordinateSystem
footprintWKT
verbatimSRS
georeferencedBy
georeferenceProtocol
georeferencedDate
georeferenceSources
georeferenceRemarks
locationAccordingTo
locationRemarks

identifiers

geographic data

vertical components

georeference

georeference metadata

other data

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.**

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.**
- **Sometimes one or more among the many location descriptions of a place have a georeference.**

Question

Has someone else already georeferenced this location well enough that I can use it?

Location terms

locationID
higherGeographyID
higherGeography
islandGroup
island
waterBody
continent
country
countryCode
stateProvince
county
municipality
locality
verbatimLocality
minimumElevationInMeters
maximumElevationInMeters
minimumDepthInMeters
maximumDepthInMeters
minimumDistanceAboveSurfaceInMeters
verbatimElevation
maximumDistanceAboveSurfaceInMeters
verbatimDepth
decimalLatitude
decimalLongitude
coordinatePrecision
verbatimLatitude
verbatimLongitude
verbatimCoordinates
coordinateUncertaintyInMeters
geodeticDatum
footprintSRS
footprintSpatialFit
pointRadiusSpatialFit
verbatimCoordinateSystem
footprintWKT
verbatimSRS
georeferencedBy
georeferenceProtocol
georeferencedDate
georeferenceSources
georeferenceRemarks
locationAccordingTo
locationRemarks

identifiers

geographic data

vertical components

georeference

georeference metadata

other data

Location terms

locationID | higherGeographyID

higherGeography | islandGroup | island | waterBody

continent | country | countryCode | stateProvince

county | municipality | locality | verbatimLocality

minimumElevationInMeters | maximumElevationInMeters

minimumDepthInMeters | maximumDepthInMeters

minimumDistanceAboveSurfaceInMeters | verbatimElevation

maximumDistanceAboveSurfaceInMeters | verbatimDepth

decimalLatitude | decimalLongitude | coordinatePrecision

verbatimLatitude | verbatimLongitude | verbatimCoordinates

coordinateUncertaintyInMeters | geodeticDatum

footprintSRS | footprintSpatialFit | pointRadiusSpatialFit

verbatimCoordinateSystem | footprintWKT | verbatimSRS

georeferencedBy | georeferenceProtocol | georeferencedDate

georeferenceSources | georeferenceRemarks

locationAccordingTo | locationRemarks

identifiers

geographic data

vertical components

georeference

georeference metadata

other data

Question

Has someone else already georeferenced this location well enough that I can use it?

Fitness for Use

	Usefulness	Occurrences	%	Locations	%
Total	-	2,232,326,955		174,245,784	
Coordinates	**mappable**	2,093,146,781	**93.8**	149,565,869	**85.8**

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

GBIF.org

Fitness for Use

Usefulness	Occurrences	%	Locations	%
Total	-	2,232,326,955		174,245,784	
Coordinates	**mappable**	2,093,146,781	**93.8**	149,565,869	**85.8**
Coordinates + uncertainty	**mappable with circle**	717,870,489	**32.2**	91,491,38	**48**

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

GBIF.org

Fitness for Use

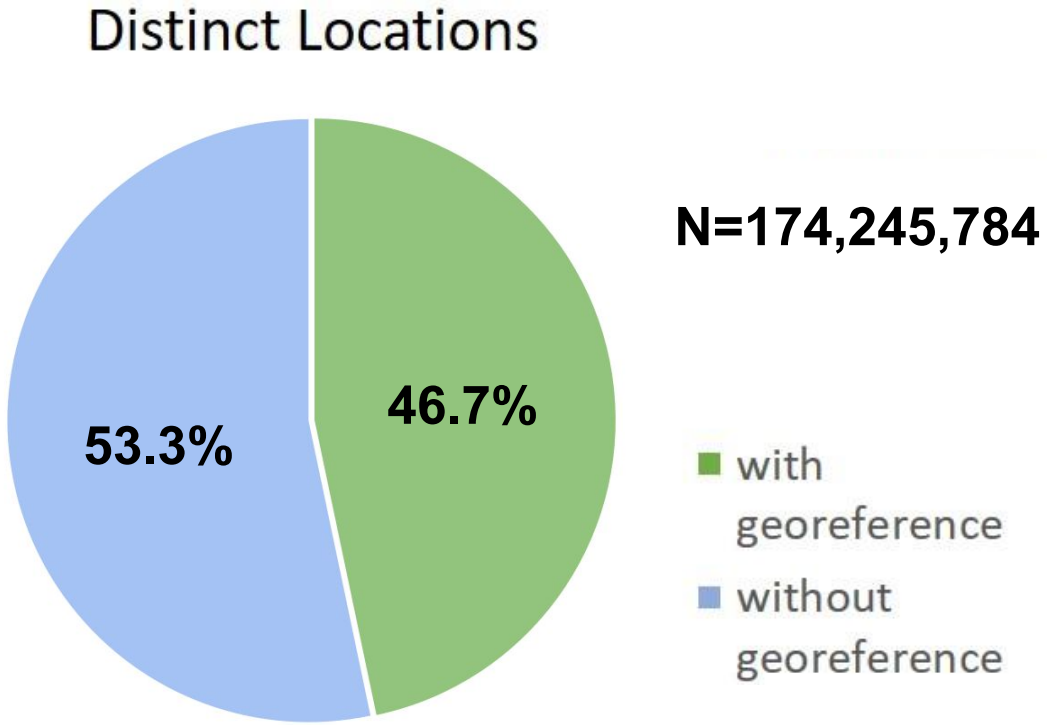
	Usefulness	Occurrences	%	Locations	%
Total	-	2,232,326,955		174,245,784	
Coordinates	**mappable**	2,093,146,781	**93.8**	149,565,869	**85.8**
Coordinates + uncertainty	**mappable with circle**	717,870,489	**32.2**	91,491,38	**48**
Coordinates + uncertainty + identifiable datum	**minimally complete**	597,054,795	**26.7**	81,320,951	**46.7**

Chapman AD and Wieczorek J. 2020. *Georeferencing Best Practices*. Copenhagen: GBIF Secretariat. https://doi.org/10.15468/doc-gg7h-s853

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

GBIF.org

Fitness for Use



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

GBIF.org

Fitness for Use

	Usefulness	Occurrences	%	Locations	%
Total	-	2,232,326,955		174,245,784	
Coordinates	**mappable**	2,093,146,781	**93.8**	149,565,869	**85.8**
Coordinates + uncertainty	**mappable with circle**	717,870,489	**32.2**	91,491,38	**48**
Coordinates + uncertainty + identifiable datum	**minimally complete**	597,054,795	**26.7**	81,320,951	**46.7**
Coordinates + Uncertainty + identifiable datum + source + protocol	**theoretically reproducible**	9,095,539	**0.41**	1,465,243	**0.84**

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?

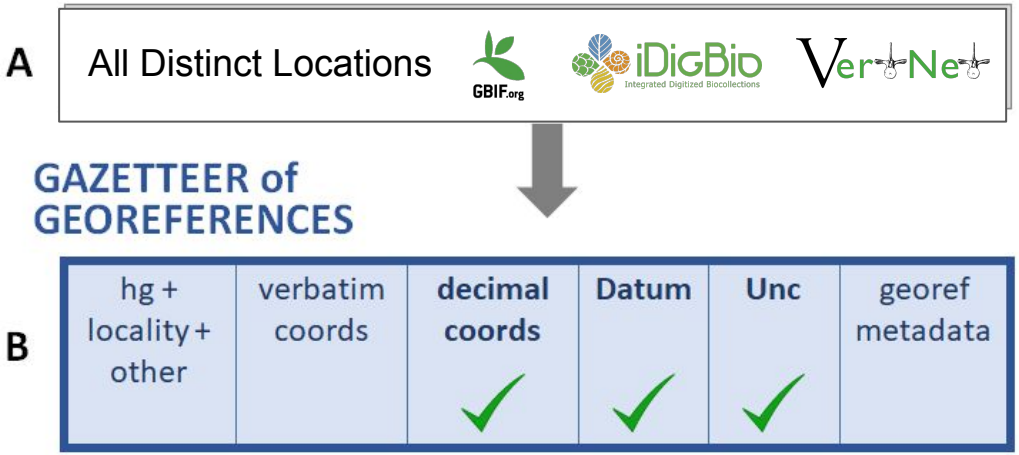
- Gazetteer of shared Locations
  - Process for matching strings
  - Compute best georeference

How does it work?

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

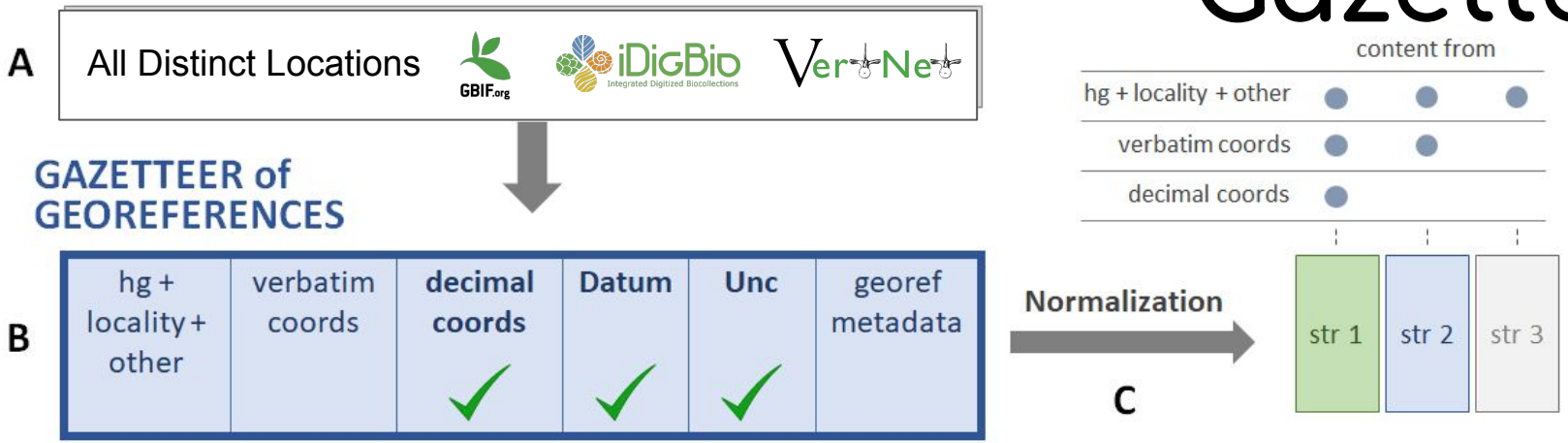
- Georeference Matcher
  - Process for matching strings
  - Find best georeference

Gazetteer



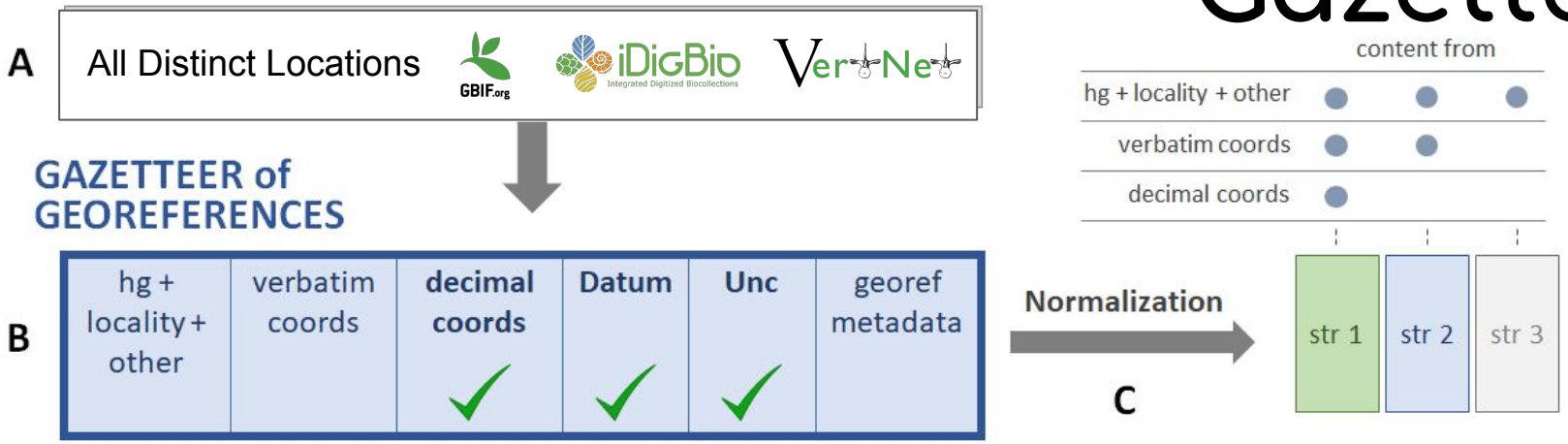
- Import Occurrences into Google BigQuery
- Assign unique identifier based on Location term contents
- Standardize countryCode
- Select for valid coordinates
- Select for valid coordinate uncertainty
- Standardize coordinate precision
- Interpret geodeticDatum
- Calculate georeference score

Gazetteer



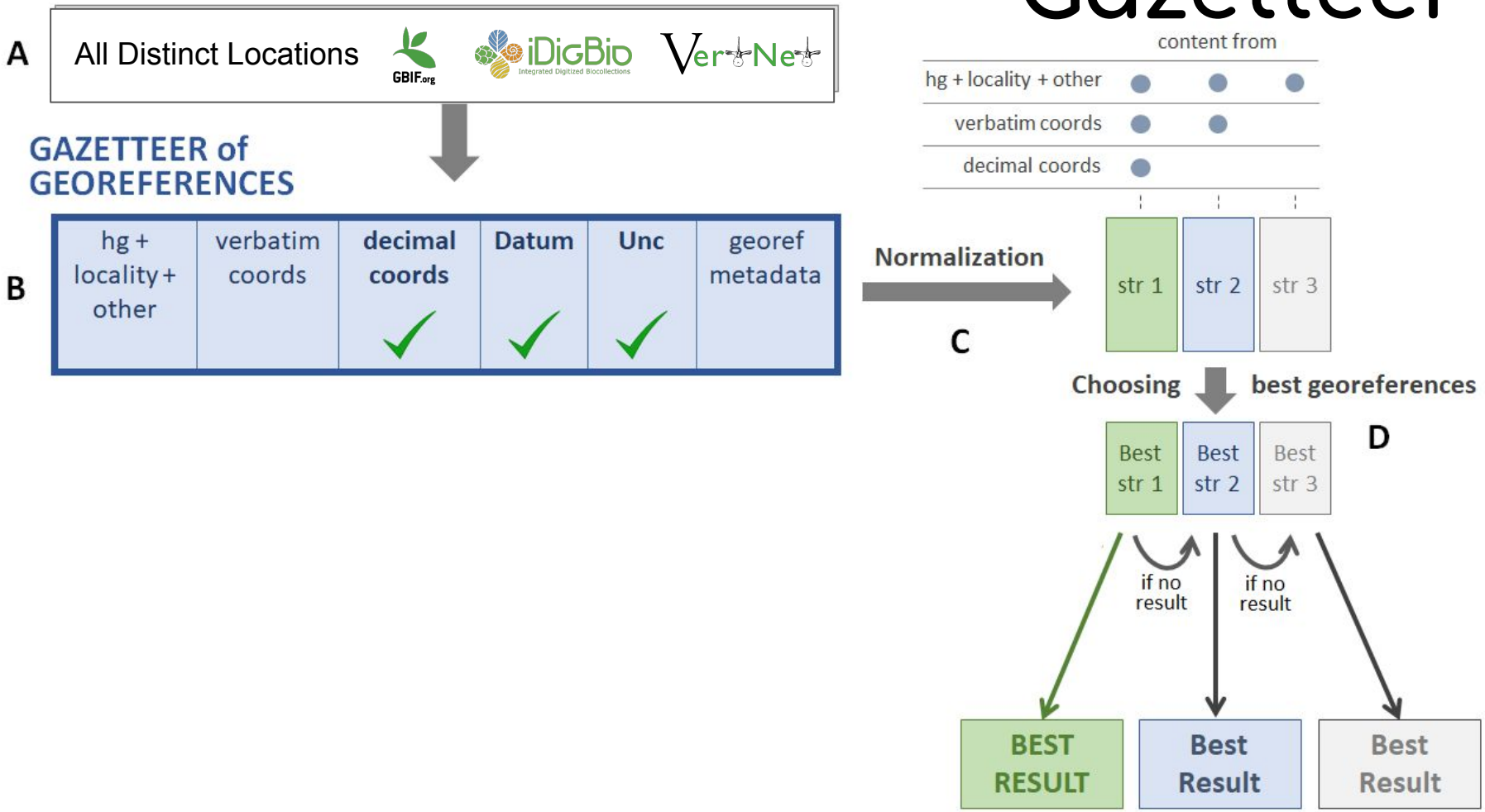
- Matching str 3 includes: higher geography (sans continent country, with interpreted countryCode), locality (collapse with verbatimLocality), elevation, and depth
- Matching str 2 includes str 3 plus verbatim coordinate terms
- Matching str 1 includes str 2 plus decimal coordinate terms

Gazetteer



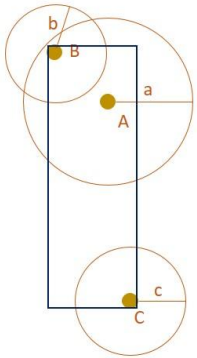
- removeSymbols(): Remove punctuation and symbols except . , / - and +
- saveNumbers(): Replace , . / - and + with space except between digits
- simplifyDiacritics(): Normalize unicode, remove white space, lowercase, and change diacritics to ASCII "equivalents"

Gazetteer


A
All Distinct Locations
GBIF.org
iDigBio
Integrated Digitized Biocollections
VerNet
GAZETTEER of GEOREFERENCES
B
hg + locality + other
verbatim coords
decimal coords
Datum
Unc
georef metadata
Normalization
C
content from
hg + locality + other
verbatim coords
decimal coords
str 1
str 2
str 3
Choosing best georeferences
D
Best str 1
Best str 2
Best str 3
if no result
if no result
BEST RESULT
Best Result
Best Result

Find the best georeference

A

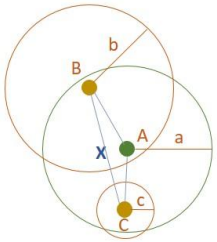


Max uncert = a → keep candidate A

Bounding box N-S > 2a

=> DISCARD

B



Max uncert = a = b → keep candidates A, B

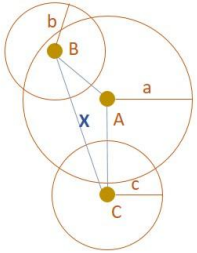
Bounding box N-S, E-W ≤ 2a, 2b → keep candidates A, B

$\overline{AX} < \overline{BX}$ → keep candidate A

$a \geq \overline{AB}, \overline{AC}$

=> KEEP A

C



Max uncert = a → keep candidate A

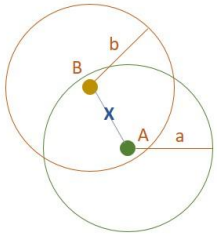
Bounding box N-S, E-W ≤ 2a → keep candidate A

$\overline{AX}$ = mín → keep candidate A

$a < \overline{AC}$

=> 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

$a \geq \overline{AB}$, $b \geq \overline{BA}$ → keep candidates A, B

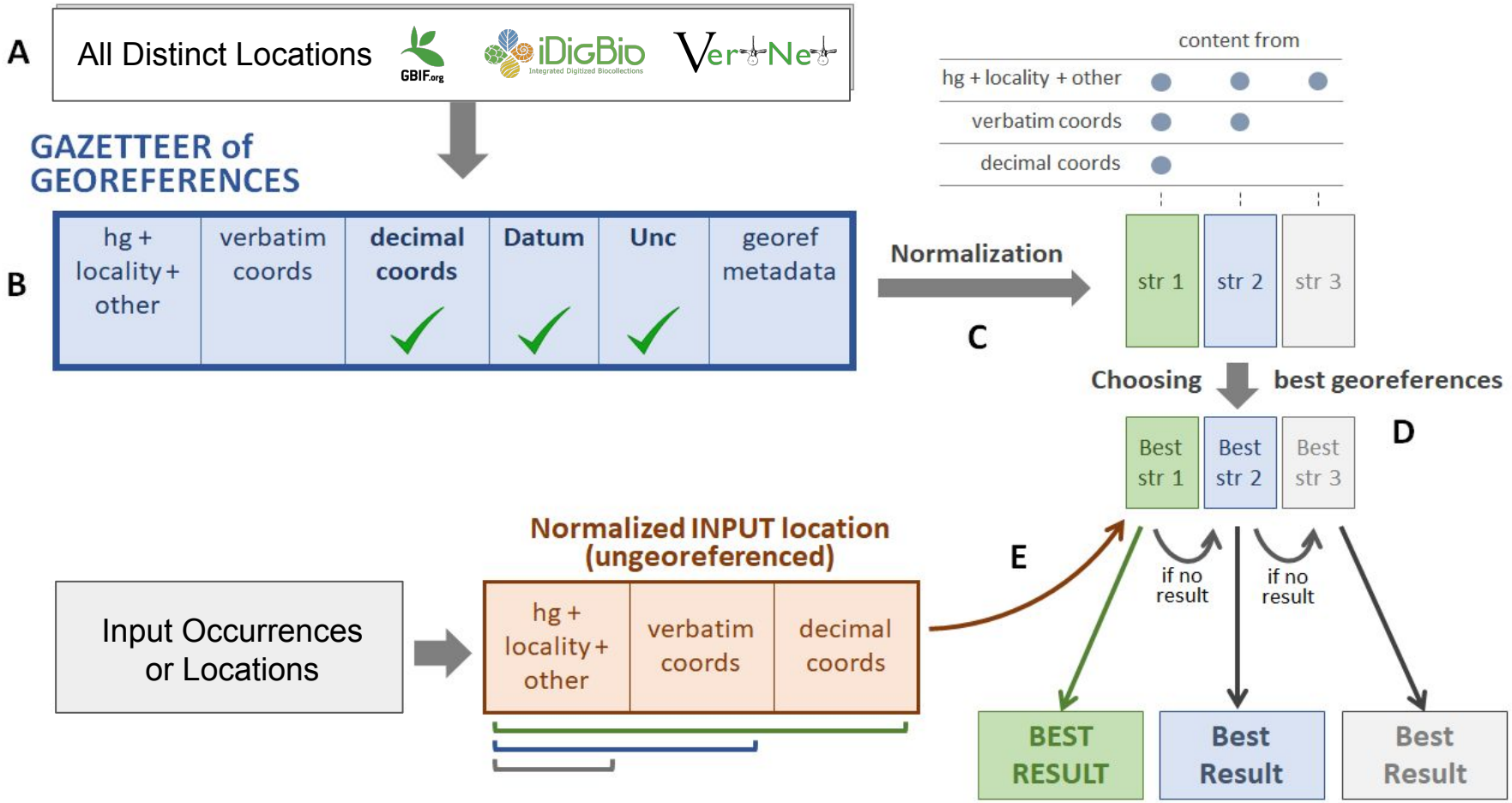
=> KEEP result with best metadata

Find the best georeference

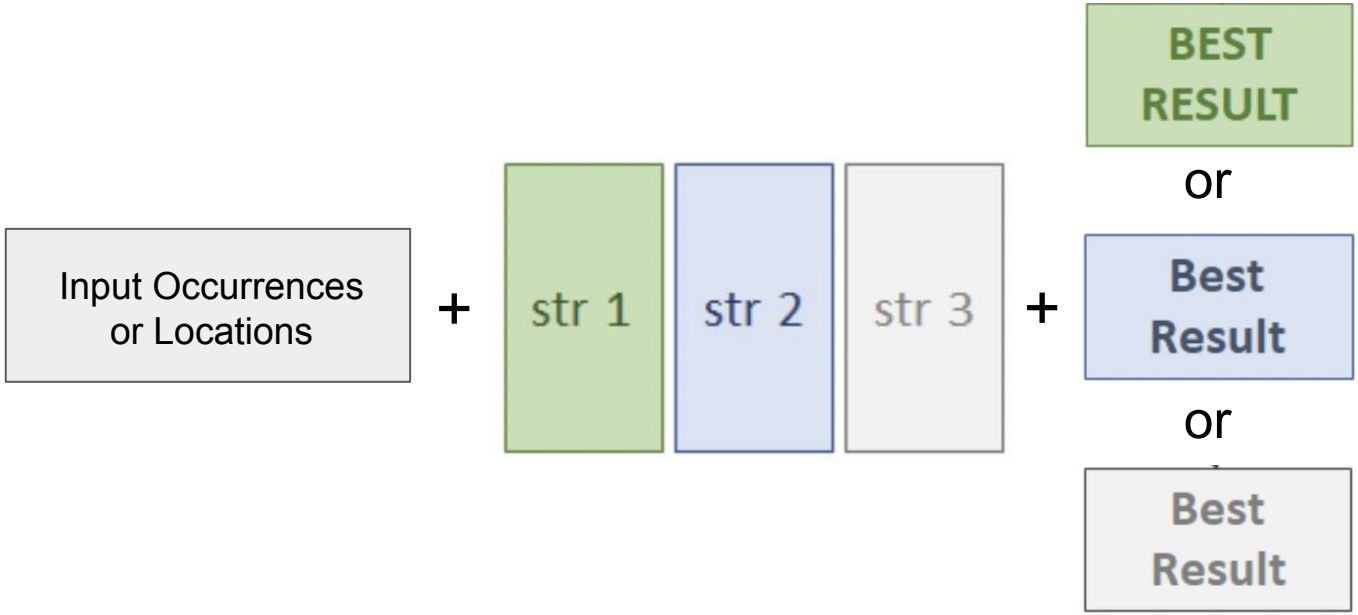
For any given string, in order:

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



Georeference Matcher Output



Best Result fields: bels_match_country, bels_interpreted_countrycode, bels_matchwithcoords, bels_matchverbatimcoords, bels_matchsanscoords, 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**

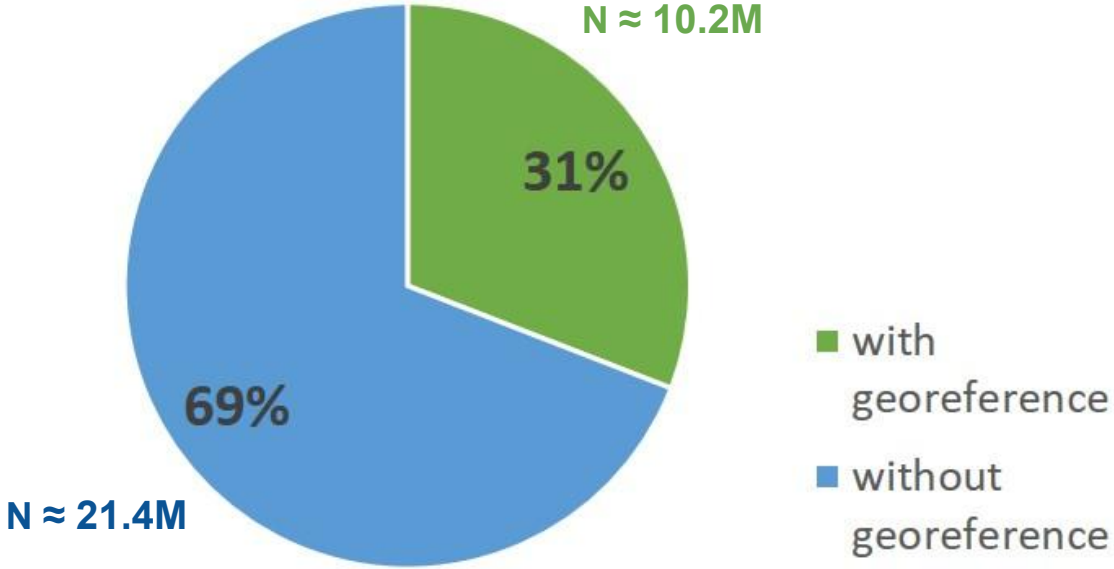
Proof of concept

- **Matching:**
  - **best practice georeferences**
  - **exact match only**
  - **no extra tricks**

Proof of concept

occurrences N-fixing species

Before

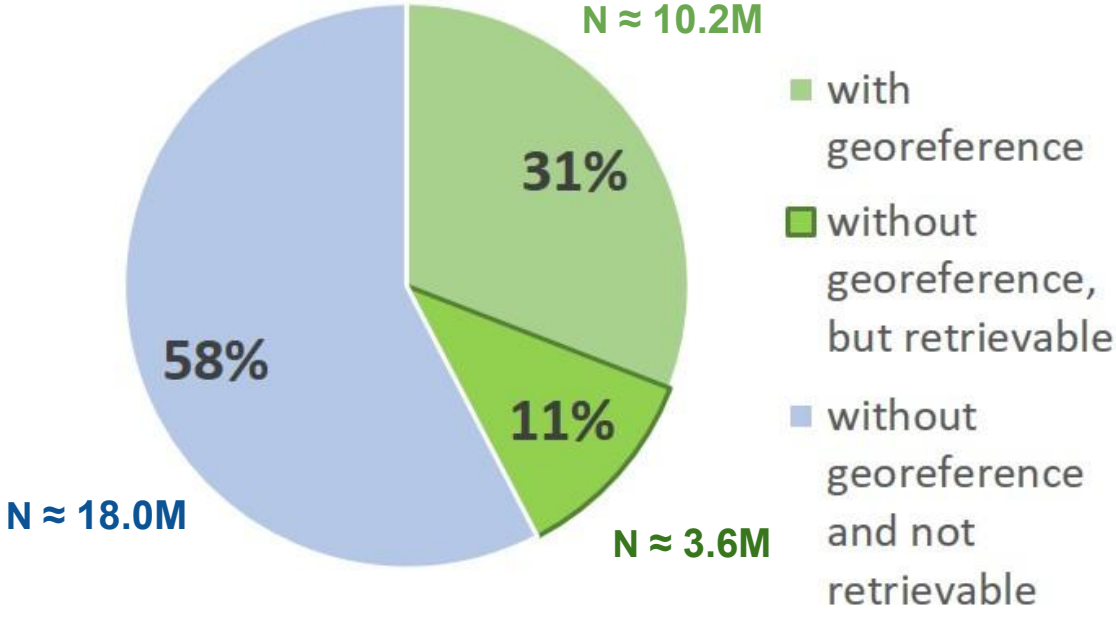


Data from GBIF snapshot 2019-04-15.

Proof of concept

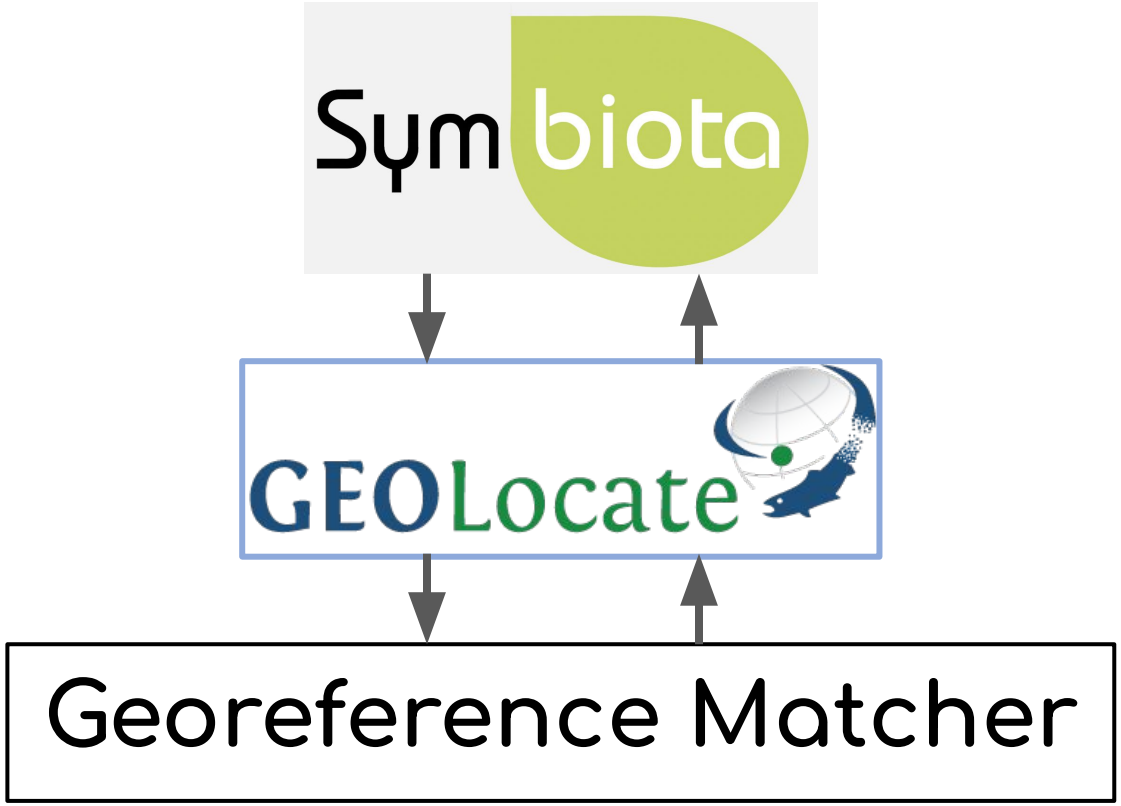
occurrences N-fixing species

After



Data from GBIF snapshot 2019-04-15.

API Integration



The BELS Georeference Matcher

Thank you!

Digi-Leap

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