

Identification of geographical free text information via Openrefine

Dominik Röpert, David Fichtmüller, Marcus Ernst

ISTC Meeting 2021-04-26

Goal

- Annotate location information with their GeoNames IDs
 - As many as possible
 - With as little effort as possible
- Project served as an exercise to get a feeling for the possibility of annotating our locality data.
- It's NOT a tool for automated large scale georeferencing so far

Situation

- BGBM Herbarium database
- > 400 000 free text entries about location
- 3 different columns
 - Location *(Fundort)*
 - LocationAndEcology *(FundortUndOeko)*
 - Remarks *(Anmerkungen)*
- Information in various languages: German, English, Spanish, Latin
- Additional Information
 - ISO Code
 - Coordinates

Approach



- Using OpenRefine
- Combination of manual curation and automated processing
- Start with the Low-Hanging Fruits
 - Most common places and formats first
- One annotation per record
 - the most precise one mentioned by name

Processing

- Export relevant columns from Herbarium DB
 - id, location, coordinates, etc
- Import in OpenRefine
 - Remove unnecessary whitespaces
- Filter: remove flagged rows
 - Rows with no or ambiguous geoinformation
- Filter: remove starred rows
 - Rows that are already processed

Processing

- Duplicate all of the columns that will be edited
 - *Location_original* and *Location_edited*
- Facet by ISO Country
- Facet by Location
 - And the other two columns (Flags, Stars)



Facet / Filter

Undo / Redo 1935 / 1935

Refresh

Reset All

Remove All

Starred Rows change invert reset

2 choices Sort by: **name** count

false 37569 exclude

true 136475

Facet by choice counts

Flagged Rows change invert reset

1 choices Sort by: **name** count

false 37569 exclude

Facet by choice counts

fkIsoCode change invert reset

233 choices Sort by: **name** count

Cluster

AD 16

AE 6

Fundort change

5264 choices total, too many to display
Set choice count limit

Facet by choice counts

37569 matching rows (414500 total)

Show as: **rows** records Show: 5 10 25 50 rows

mm	Sammeldatum	Anmerkungen	HerbariumID	StableURI	Co
	2012-06-19	Herbariumnummer in Berlin: B 10 0502749; utm: EJ 69.47.06	B100502749	http://herbarium.bgbm.org/object/B100502749	2012-0
	2012-06-19	Herbariumnummer in Berlin: B 10 0499603; utm: EJ 69.47.06	B100499603	http://herbarium.bgbm.org/object/B100499603	2012-0
	2012-06-19	Herbariumnummer in Berlin: B 10 0502748; utm: EJ 69.47.06	B100502748	http://herbarium.bgbm.org/object/B100502748	2012-0
	2012-06-15	Herbariumnummer in Berlin: B 10 0499611; utm: EK 51.15.43	B100499611	http://herbarium.bgbm.org/object/B100499611	2012-0
	2012-06-08	Herbariumnummer in Berlin: B 10 0497517; utm: EK 59,86.48	B100497517	http://herbarium.bgbm.org/object/B100497517	2012-0
	2012-06-05	Herbariumnummer in Berlin: B 10 0493695; utm: EK 69.74.68	B100493695	http://herbarium.bgbm.org/object/B100493695	2012-0
	2010-05-30	Herbariumnummer in Berlin: B 10 0405038; utm: EH 83.32.93	B100405038	http://herbarium.bgbm.org/object/B100405038	2010-0
	2010-05-29	Herbariumnummer in Berlin: B 10 0405040; utm: FH 21.25.50	B100405040	http://herbarium.bgbm.org/object/B100405040	2010-0
	2009-04-28	Herbariumnummer in Berlin: B 10 0201733	B100291733	http://herbarium.bgbm.org/object/B100291733	2009-0

Clustering

- Free text information of many entries vary often slightly
- Identify and cluster them

Clustering

Cluster & Edit column "Fundort"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. [Find out more...](#)

Method **key collision** Keying Function **cologne-phonetic** **4760 clusters found**

4	6	<ul style="list-style-type: none">Sérrai, Sintikis, 2,7 km W Fea Petra (3 rows)Sérrai, Sintikis, 1,5 km W Fea Petra (4 rows)Sérrai, Sintikis, 1,2 km N Fea Petra (2 rows)Sérrai, Sintikis, 1,8 km NNO Fea Petra (1 rows)	<input type="checkbox"/>	Sérrai, Sintikis, 2,7 km W Fea Petra
4	6	<ul style="list-style-type: none">Magnisia, Almyrou, 11,6 km SO Almiros (2 rows)Magnisia, Almyrou, 3,4 km SO Almiros (2 rows)Magnisia, Almyrou, 2,8 km SO Almiros (1 rows)Magnisia, Almyrou, 6,8 km SO Almiros (1 rows)	<input type="checkbox"/>	Magnisia, Almyrou, 11,6 km SO A
4	244	<ul style="list-style-type: none">Kreta (218 rows)Crete (19 rows)Crete (6 rows)Krete (1 rows)	<input checked="" type="checkbox"/>	Kreta
4	5	<ul style="list-style-type: none">Magnisia, Volou, 3,0 km SO Portaria (2 rows)Magnisia, Volou, 2,0 km SO Portaria (1 rows)Magnisia, Volou, 2,6 km SO Portaria (1 rows)Magnisia, Volou, 3,3 km SO Portaria (1 rows)	<input type="checkbox"/>	Magnisia, Volou, 3,0 km SO Porta
4	46	<ul style="list-style-type: none">cult. Hort. bot. Berol. (43 rows)cult. Hort. bot. Berol., 168-06-85-20 (1 rows)cult. Hort. bot. Berol., 168-07-85-20. (1 rows)cult. Hort. bot. Berol., 313-04-87-20 (1 rows)	<input type="checkbox"/>	cult. Hort. bot. Berol.

Choices in Cluster
1 — 10

Rows in Cluster
0 — 340

Average Length of Choices
0 — 230

Length Variance of Choices
0 — 6.5

Select All Unselect All Export Clusters Merge Selected & Re-Cluster Merge Selected & Close Close

Detect, Select and Process

- Look for common patterns
- „Magnisia, Almyrou, 6,8 km SO Almiros“
 - Search and Replace with RegEx
 - `(\w+,)(\w+,)?\d(,\d)? km [NWSOE]+ (\w+)`
 - \$1\$2\$4
 - => „Magnisia, Almyrou, Almiros“
- Create another column „Region“ with the first segment

Querying GeoNames

- 2 options to query
 - OpenRefine function: Add column by fetching urls
 - custom Python code
- API Limits:
 - Username required
 - 1000 requests per hour
 - 20 000 requests per day
- Store results for identical queries
- Fuzzy Matching is supported

Querying GeoNames

```
http://api.geonames.org/searchJSON  
?name=<location>  
&country=<countryISO>  
&adminCode2=<Code_of_ADM_Region>  
&fuzzy=0.6  
&username=<username>
```

Querying GeoNames

▼ geonames:

▼ 0:

```
adminCode1: "16"  
lng: "13.41053"  
geonameId: 2950159  
toponymName: "Berlin"  
countryId: "2921044"  
fcl: "P"  
population: 3426354  
countryCode: "DE"  
name: "Berlin"  
fclName: "city, village,..."
```

▼ adminCodes1:

```
ISO3166_2: "BE"  
countryName: "Germany"  
fcodeName: "capital of a political entity"  
adminName1: "Berlin"  
lat: "52.52437"  
fcode: "PPLC"
```

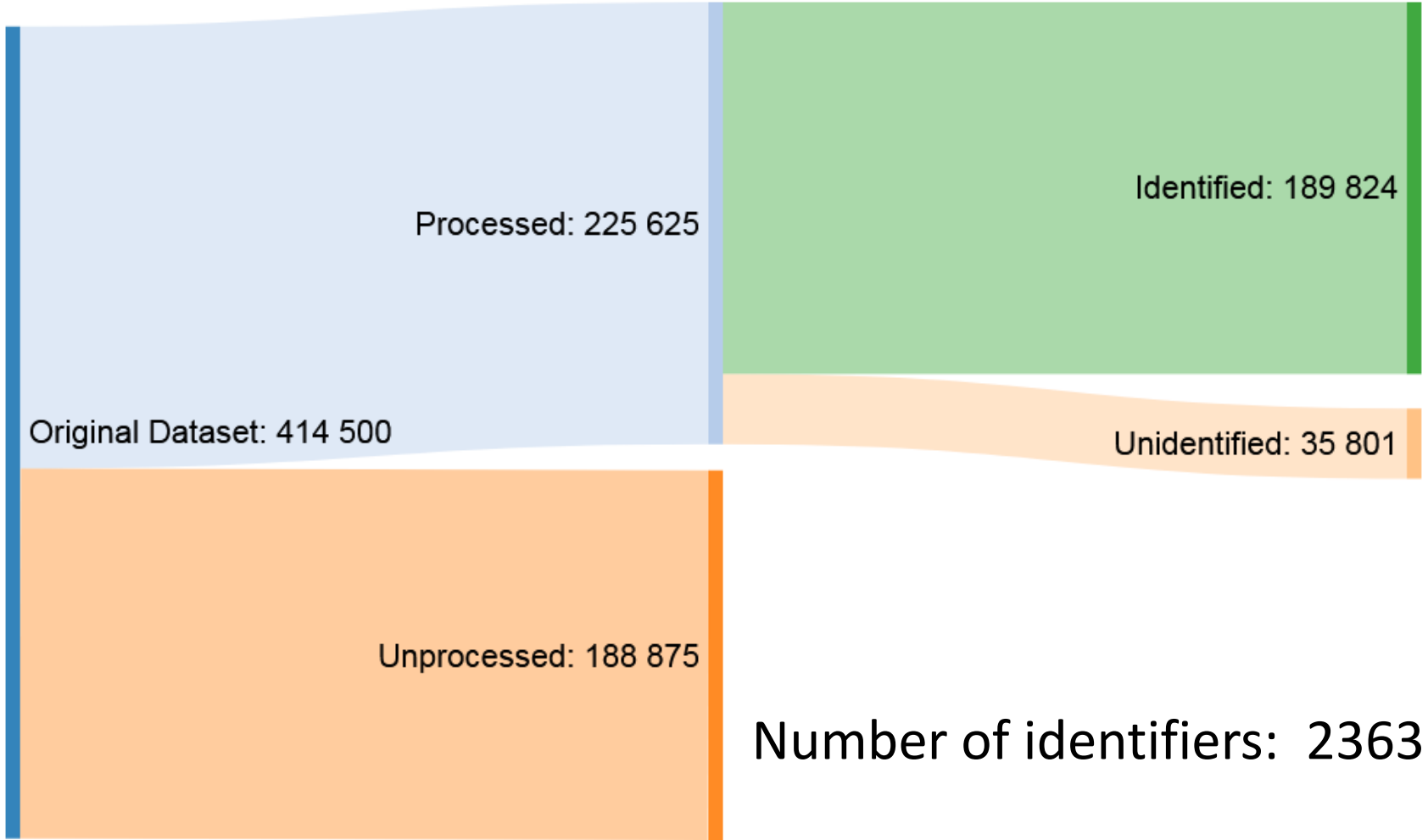
Validating Results

- Calculate distance between coordinates and GeoName result
- Warning if distance is above a certain threshold

Publishing Results in rdf

```
<rdf:RDF>
  <!--This is metadata about this metadata document-->
  <rdf:Description rdf:about="http://herbarium.bgbm.org/data/rdf/B100327891">
    <dc:subject rdf:resource="http://herbarium.bgbm.org/object/B100327891"/>
    <dc:created>2021-04-25T21:20:17+02:00</dc:created>
  </rdf:Description>
  <!--This is metadata about this specimen-->
  <rdf:Description rdf:about="http://herbarium.bgbm.org/object/B100327891">
    <dc:title>Veronica teucrium L.</dc:title>
    <dc:description>
      A herbarium specimen of Veronica teucrium L. collected by Billot,C.
    </dc:description>
    <dc:creator>Billot,C.</dc:creator>
    <dc:created>1854-06-11</dc:created>
    <dc:type>Specimen</dc:type>
    <dc:publisher>BGBM</dc:publisher>
    <dwc:materialSampleID>https://herbarium.bgbm.org/object/B100327891</dwc:materialSampleID>
    <dwc:basisOfRecord>Specimen</dwc:basisOfRecord>
    <dwc:collectionCode>B</dwc:collectionCode>
    <dwc:catalogNumber>B 10 0327891</dwc:catalogNumber>
    <dwc:scientificName>Veronica teucrium L.</dwc:scientificName>
    <dwc:previousIdentifications>Veronica teucrium L.</dwc:previousIdentifications>
    <dwc:family>Scrophulariaceae</dwc:family>
    <dwc:genus>Veronica</dwc:genus>
    <dwc:specificEpithet>teucrium</dwc:specificEpithet>
    <dwc:country>Germany</dwc:country>
    <dwc:countryCode>DE</dwc:countryCode>
    <dwc:locality>Berlin</dwc:locality>
    <dwc:eventDate>1854-06-11</dwc:eventDate>
    <dwc:recordNumber>1729</dwc:recordNumber>
    <dwc:recordedBy>Billot,C.</dwc:recordedBy>
    <dwc:fieldNumber>1729</dwc:fieldNumber>
    <dwciri:recordedBy rdf:resource="http://www.wikidata.org/entity/Q7149945"/>
    <dwciri:inDescribedPlace rdf:resource="https://sws.geonames.org/2950159"/>
  </rdf:Description>
  <rdf:Description rdf:about="http://www.wikidata.org/entity/Q7149945">
    <owl:sameAs rdf:resource="http://purl.oclc.org/net/edu.harvard.huh/guid/uuid/cda2f653-fb10-4439-a349-c2e27144c023"/>
    <owl:sameAs rdf:resource="http://viaf.org/viaf/212281451"/>
  </rdf:Description>
</rdf:RDF>
```

Statistics



Alternative Approaches

- Wikidata
 - More synonyms
 - Build-in reconciliation service in OpenRefine
 - Usually geographic items have the GeoNames id specified
- Natural Language Processing (NLP)

Thank you

Code available at:

<https://git.bgbm.org/data-cleanup/geographical-annotation>

Questions?

Freie Universität Berlin

Botanic Garden and Botanical Museum Berlin