Identification of geographical free text information via Openrefine

David Fichtmueller, Dominik Röpert, Marcus Ernst

MOBILISE Workshop 2020-02-11 Warsaw

Caveat

- I will NOT present you a tool for automated large scale georeferencing
- Data is messy and full of ambiguities
- Automated processing is occasionally possible for selected subsets
- Manual curation and oversight is always necessary
- Your millage may vary: depending on the structure of the data

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

Goal

- Annotate location information with their GeoNames IDs
  - As many as possible
  - With as little effort as possible

Approach

- Usinge 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 unnessary 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

OpenRefine Botanical Data

Permalink

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)

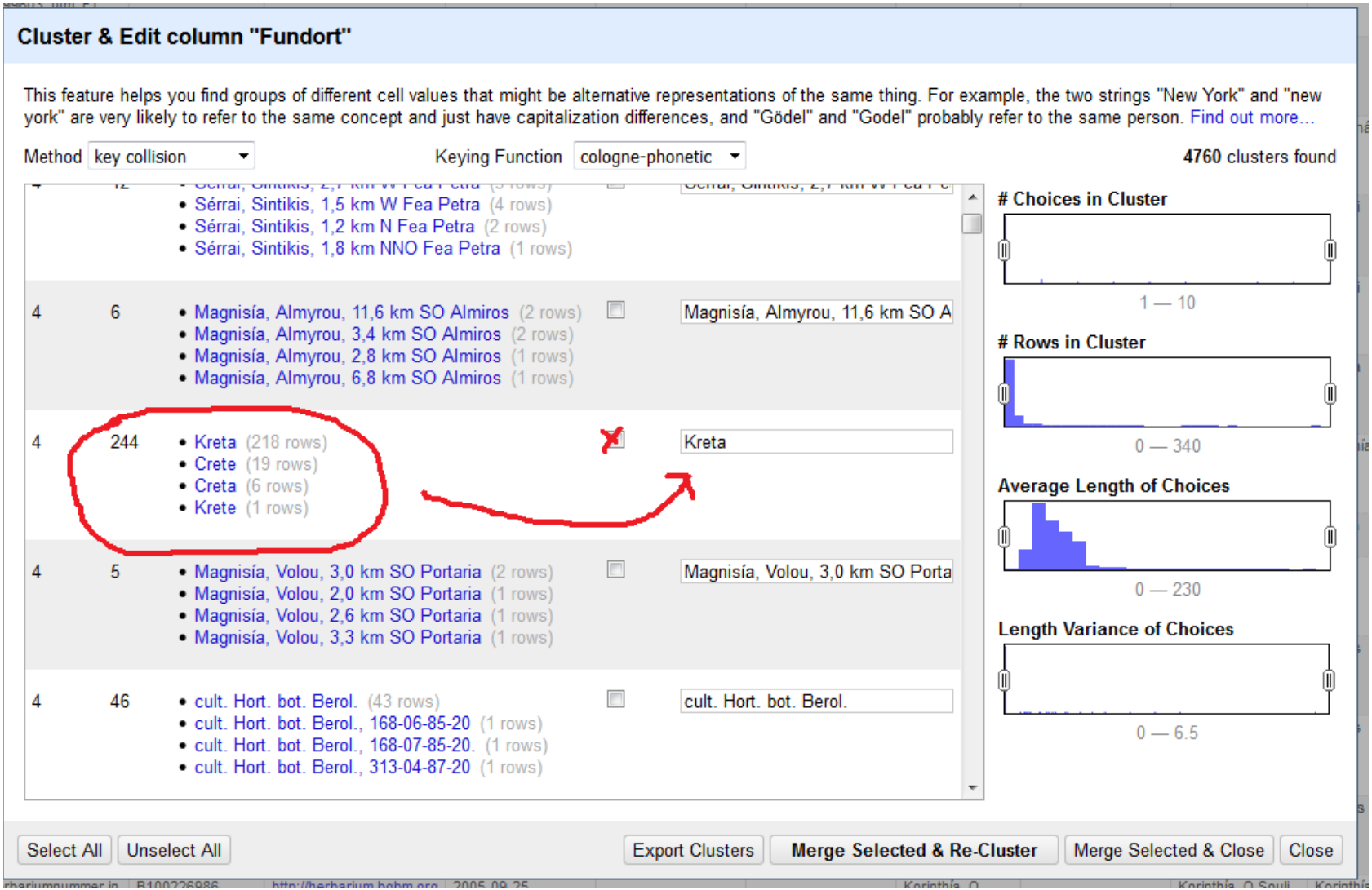
Show as: **rows** records Show: 5 10 25 **50** rows

mme	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	B100291733	http://herbarium.bgbm.org	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
• 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)
4 6
• Magnisía, Almyrou, 11,6 km SO Almiros (2 rows)
• Magnisía, Almyrou, 3,4 km SO Almiros (2 rows)
• Magnisía, Almyrou, 2,8 km SO Almiros (1 rows)
• Magnisía, Almyrou, 6,8 km SO Almiros (1 rows)
Magnisía, Almyrou, 11,6 km SO A
4 244
• Kreta (218 rows)
• Crete (19 rows)
• Creta (6 rows)
• Krete (1 rows)
Kreta
4 5
• Magnisía, Volou, 3,0 km SO Portaria (2 rows)
• Magnisía, Volou, 2,0 km SO Portaria (1 rows)
• Magnisía, Volou, 2,6 km SO Portaria (1 rows)
• Magnisía, Volou, 3,3 km SO Portaria (1 rows)
Magnisía, Volou, 3,0 km SO Porta
4 46
• 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)
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

Work on Data Subsections

- Use the ISO Country field to reduce the number of choices in the Location facet
- Disadvantage: has to be repeated for each country

Detect, Select and Process

- Look for common patterns

- „Herbarium Nr: 1234567“
  - Filter by RegEx: „^Herbarium Nr\.?: \d+$“
  - Flag rows

- „Magnisia, Almyrou, 6,8 km SO Almiros“
  - Search and Replace with RegEx
    - (\w+,)(\w+,)?\d(,\d)? km [NWSOE]+ (\w+)
    - $1$2$4

Further subsections if necessary

- „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:
  - Usename 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": [{
    – "adminCode1": "78",
    – "lng": "21.01178",
    – "geonameId": 756135,
    – "toponymName": "Warsaw",
    – "countryId": "798544",
    – "fcl": "P",
    – "population": 1702139,
    – "countryCode": "PL",
    – "name": "Warsaw",
    – "fclName": "city, village,...",
    – "adminCodes1": {
        • "ISO3166_2": "14"
    – },
    – "countryName": "Poland",
    – "fcodeName": "capital of a political entity",
    – "adminName1": "Mazovia",
    – "lat": "52.22977",
    – "fcode": "PPLC"
},
```

Validating Results

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

Star and Repeat

- Star the processed rows
- Repeat for the next country/region

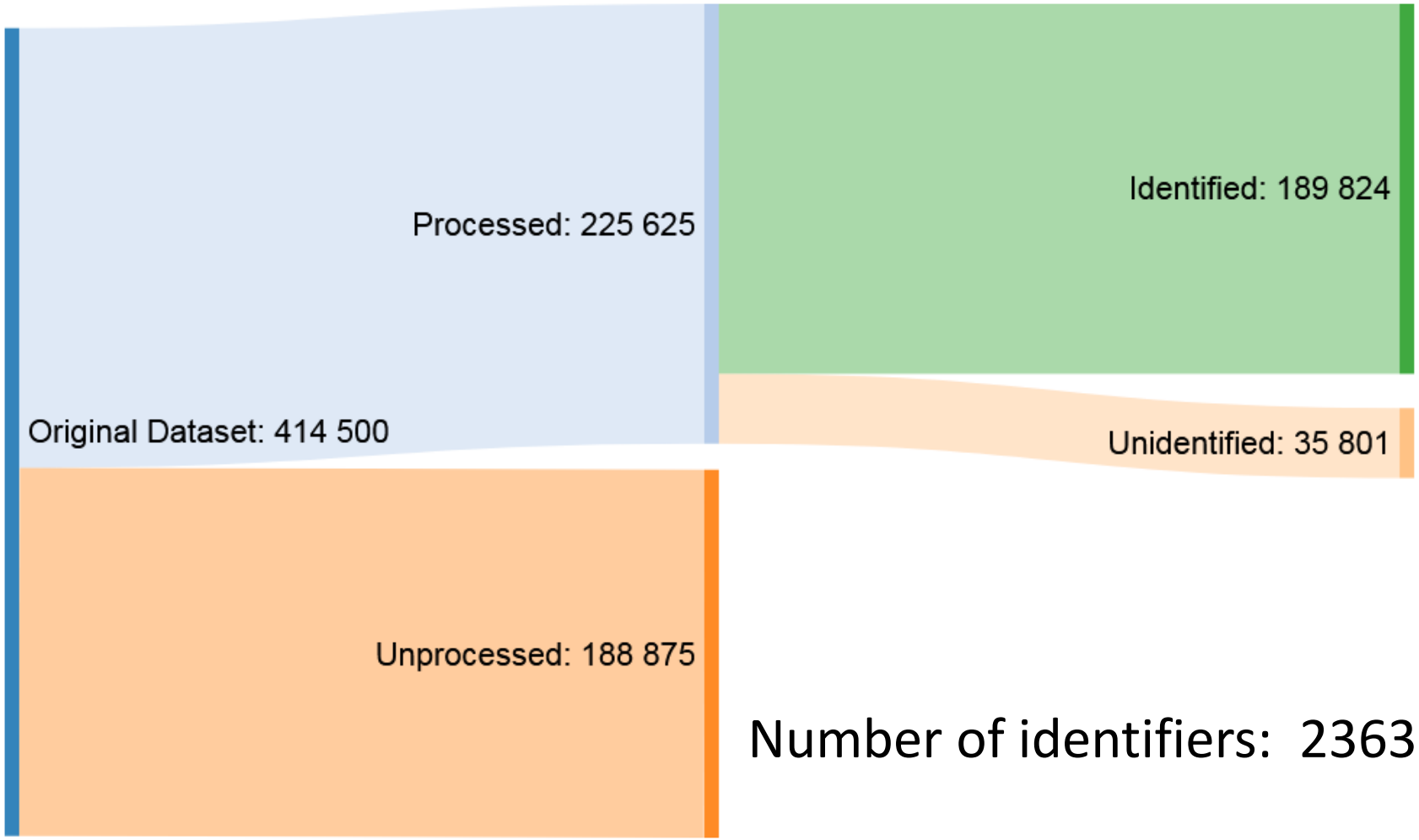
Alternative Approaches

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

Alternative Approaches

- Reverse Geocoding
  - Lookup closest place name to coodinates
  - Problem: coodinate precision
    - 31.1 is not 31.10000
    - Center of country as coodinate
- Infer Location: Collection Date & Collector

Statistics



Number of identifiers: 2363

Tale of Caution

- 7 years ago I wrote code to match country names in various languages against a location free text field
- I ran it against GBIF data and it was able to match a lot of records without country information or coordinates to a country
- A lot of the resulting matched were assigned to Iceland: Why? Do Icelandic researchers forget to specifiy the country more often than others?

Tale of Caution

- The German name for Iceland is „Island“
- It would match locations like „Alcatraz Island“

- Lession: Double check your work for plausibility

Thank you

Code available at:

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

What are your questions?

David Fichtmueller

Freie Universität Berlin

Botanic Garden and Botanical Musem Berlin

d.fichtmueller@bgbm.org