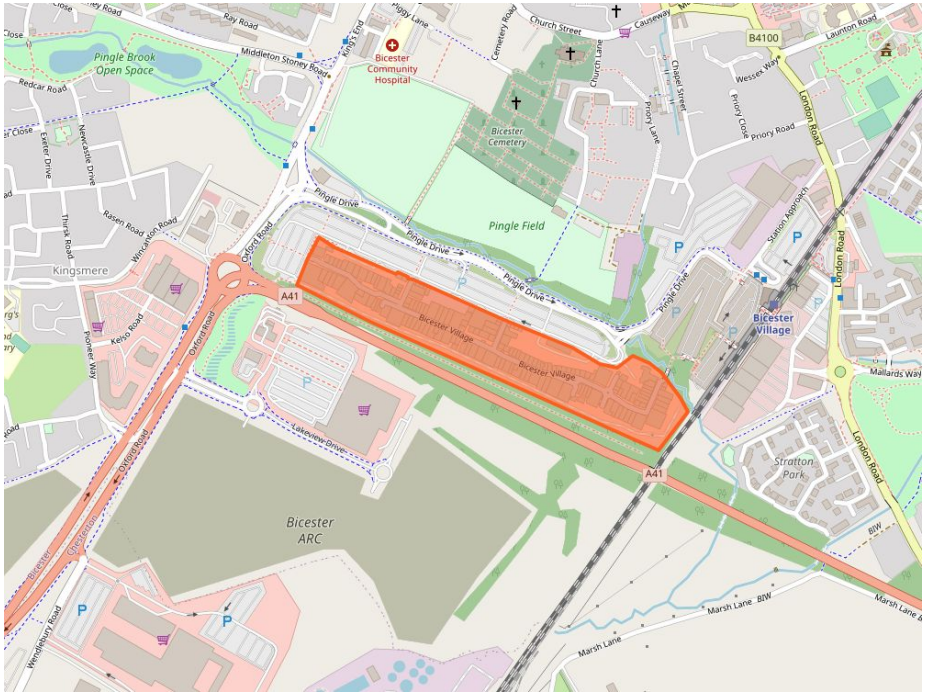
huq

SpiderMatch

Validating scraped POI data with
OpenStreetMap and VertexAI

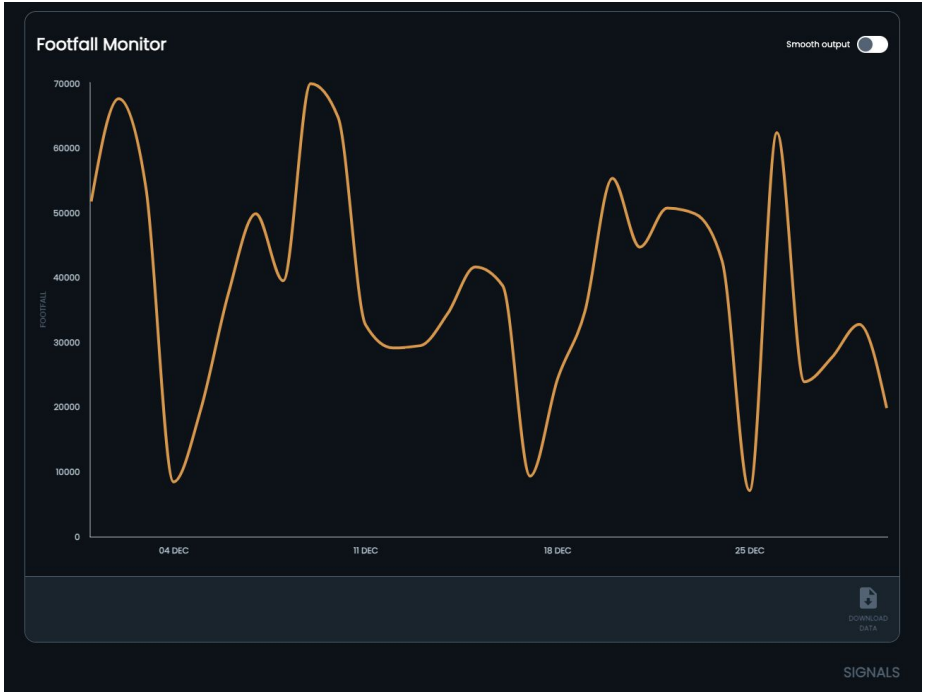
About Huq

- Modular place monitoring platform
- Clients across local government, real estate and retail



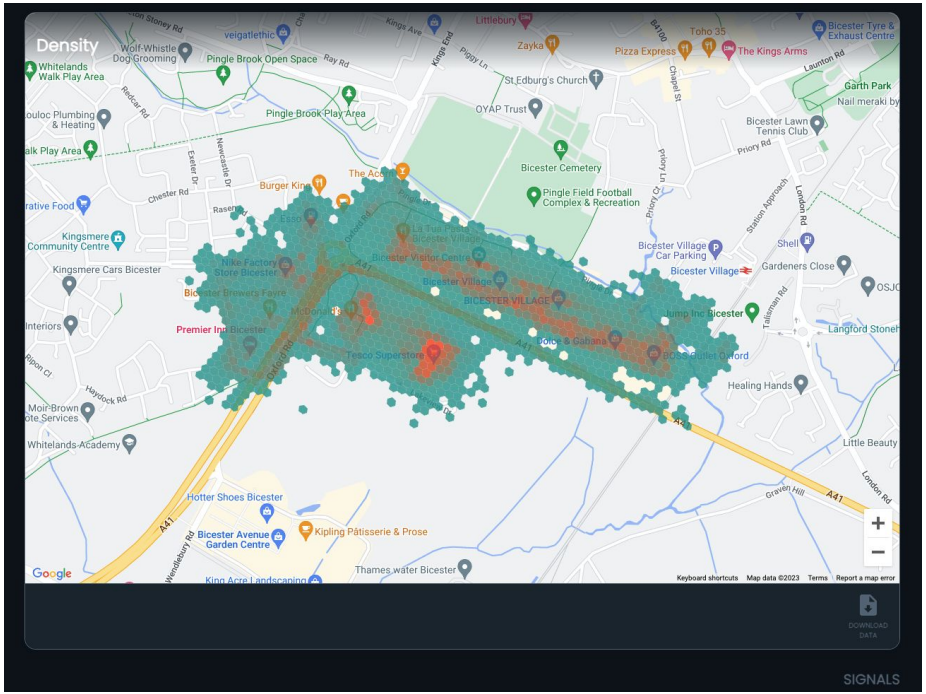
About Huq

- Modular place monitoring platform
- Clients across local government, real estate and retail
- Measures:
  - Footfall



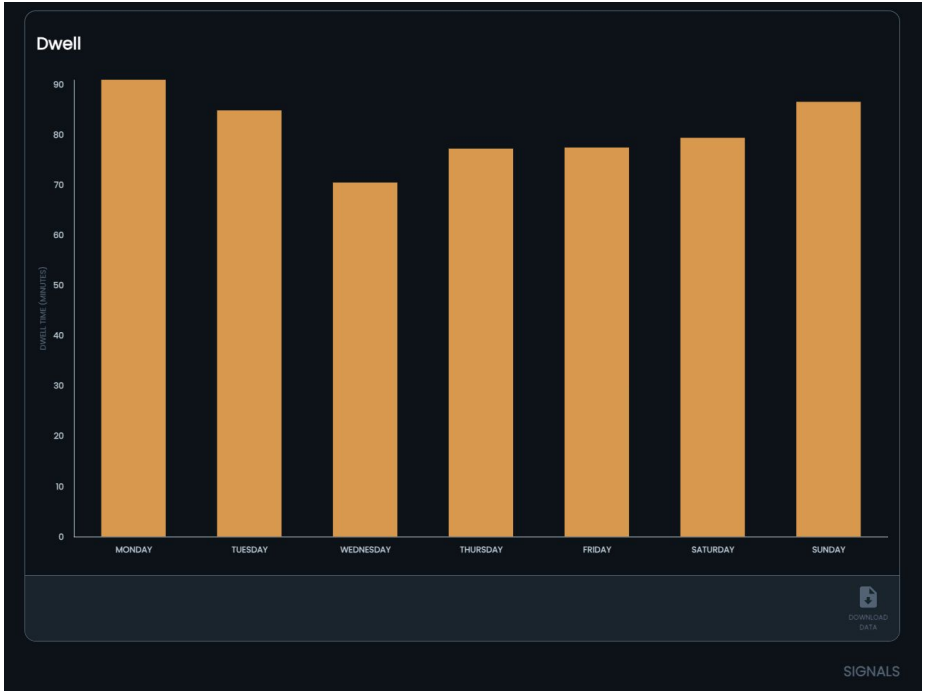
About Huq

- Modular place monitoring platform
- Clients across local government, real estate and retail
- Measures:
  - Footfall
  - Visit density



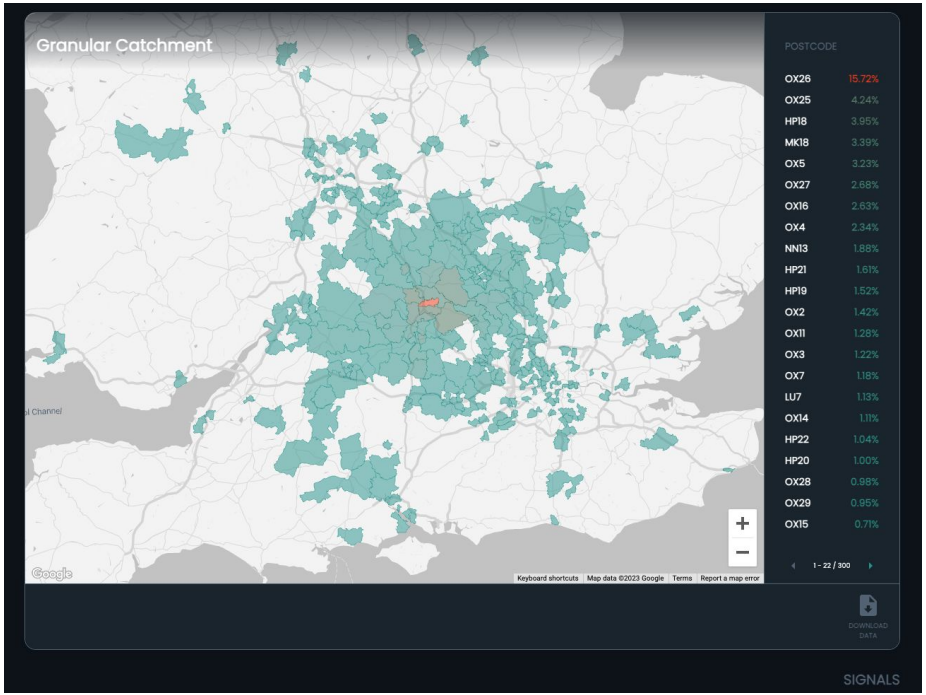
About Huq

- Modular place monitoring platform
- Clients across local government, real estate and retail
- Measures:
  - Footfall
  - Visit density
  - Dwell time



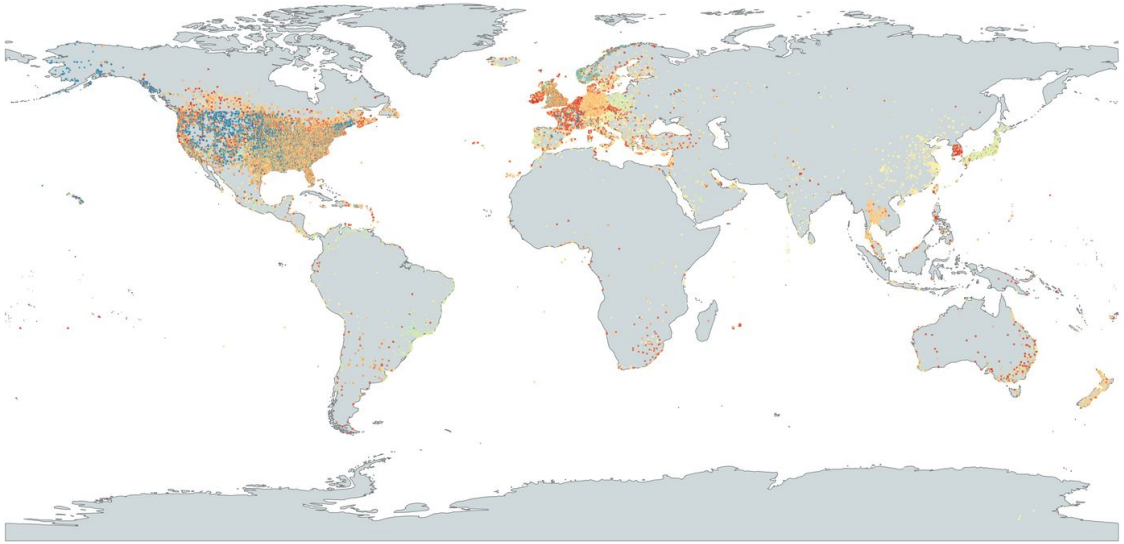
About Huq

- Modular place monitoring platform
- Clients across local government, real estate and retail
- Measures:
  - Footfall
  - Visit density
  - Dwell time
  - Granular catchment



What we scrape - AllThePlaces

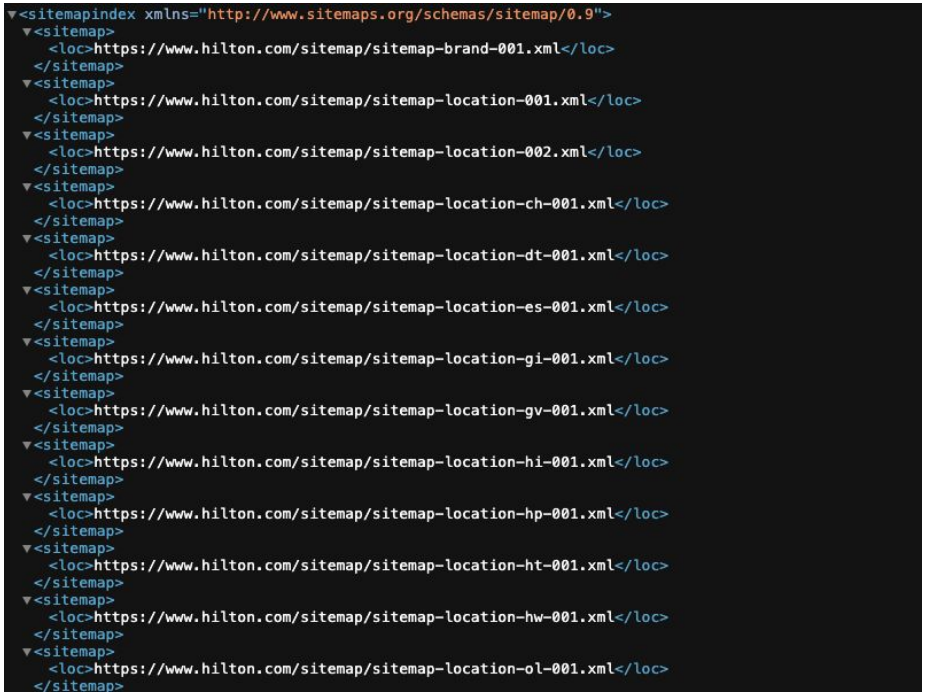
- Open-source
- ~1800 spiders
- ~4.5m items
- Runs on Scrapy Cloud
- Zyte API for proxy
- Feeds into BigQuery



++

Types of Spiders

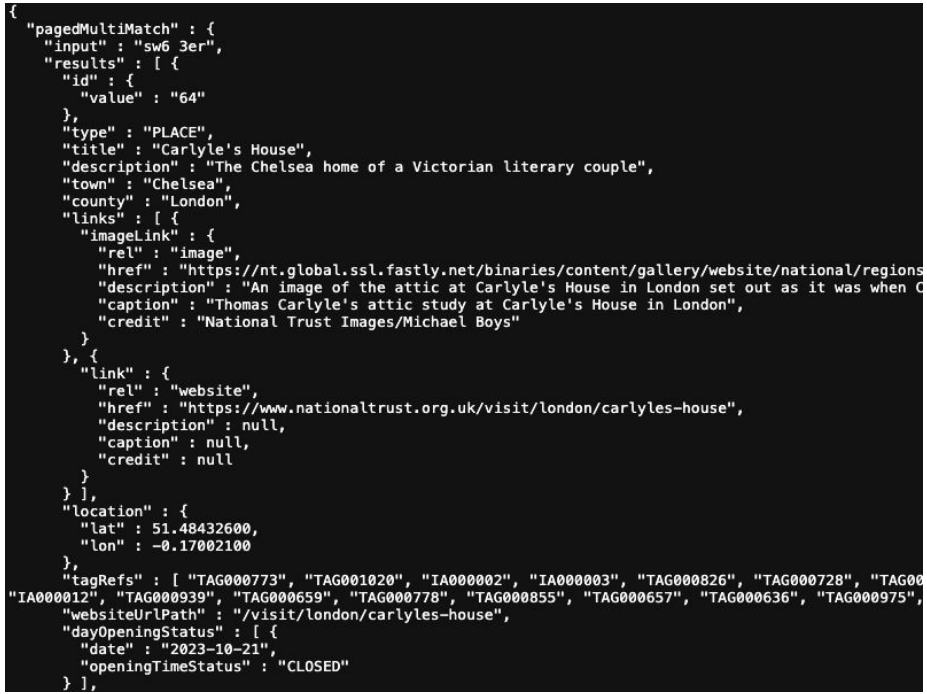
- Sitemap
- JSON API
  - Sometimes returns everything
  - Otherwise, coordinate or postcode search



++

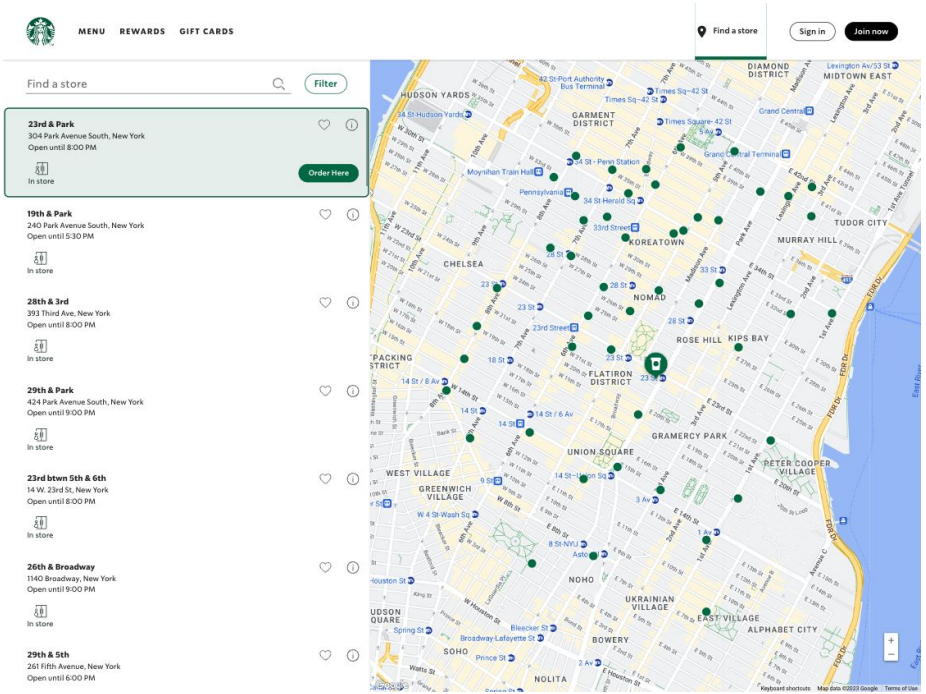
Types of Spiders

- Sitemap
- JSON API
  - Sometimes returns everything
  - Otherwise, coordinate or postcode search



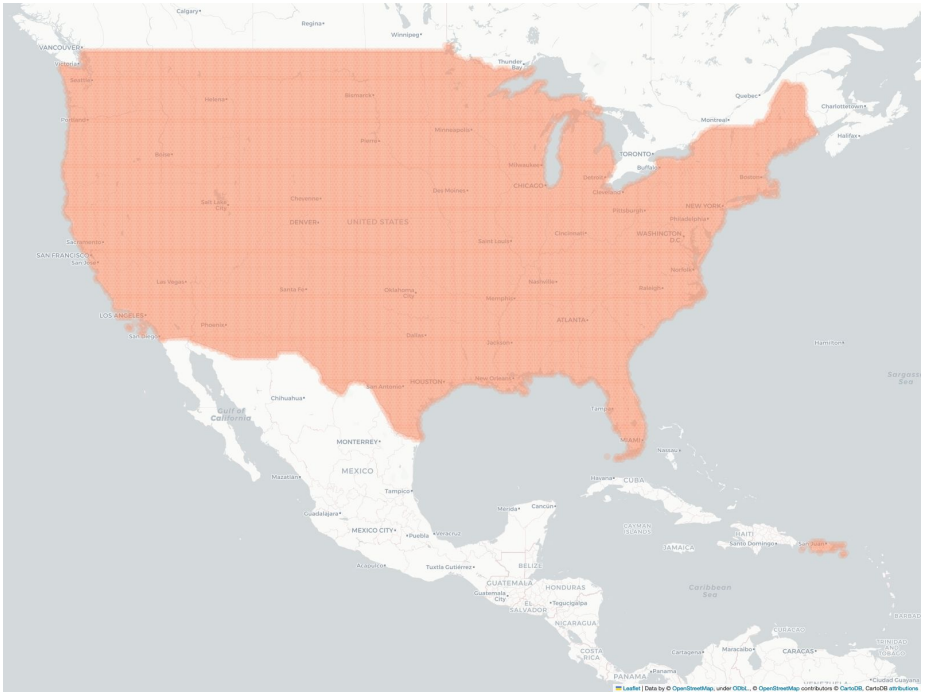
Types of Spiders

- Sitemap
- JSON API
  - Sometimes returns everything
  - Otherwise, coordinate or postcode search



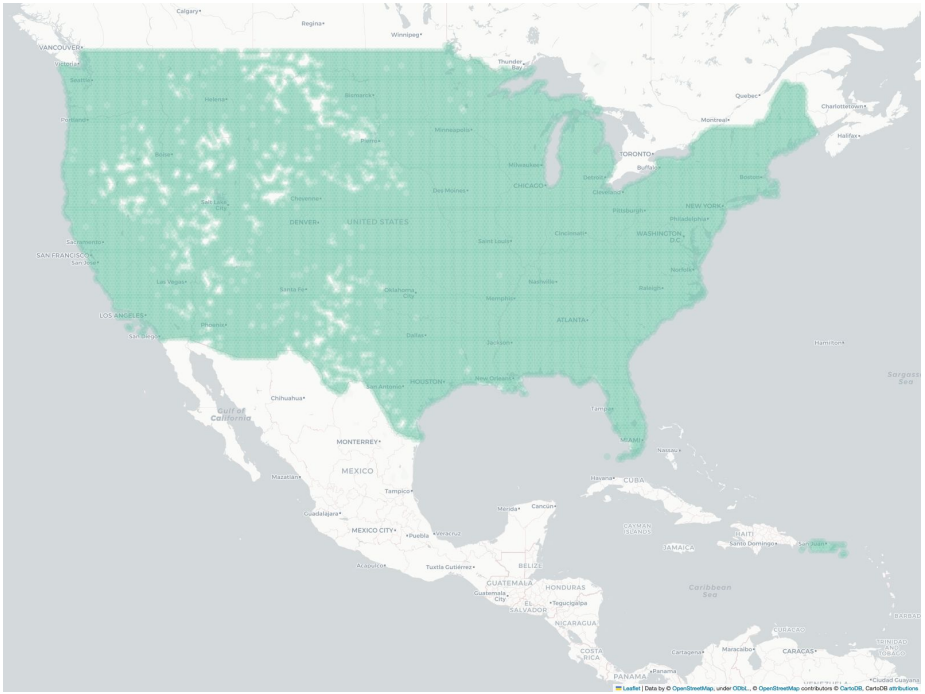
Intelligent Coordinate Searching

- 34,000 points for the US at 10mi radius
- No result pagination
- Recursive search at smaller radius when there are too many results



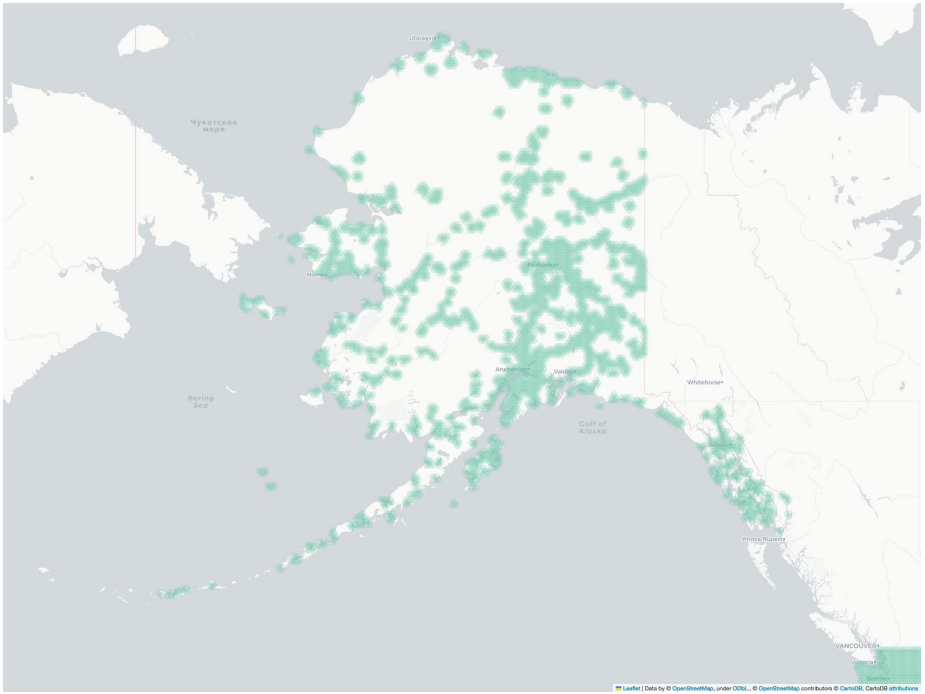
Intelligent Coordinate Searching

- 34,000 for the US at 10mi radius - so at least 34,000 requests needed
- No result pagination
- Recursive search at smaller radius when there are too many results
- What if we only searched in areas that had buildings?
    - 32% reduction in the USA
    - Now only 23,000 requests



Intelligent Coordinate Searching

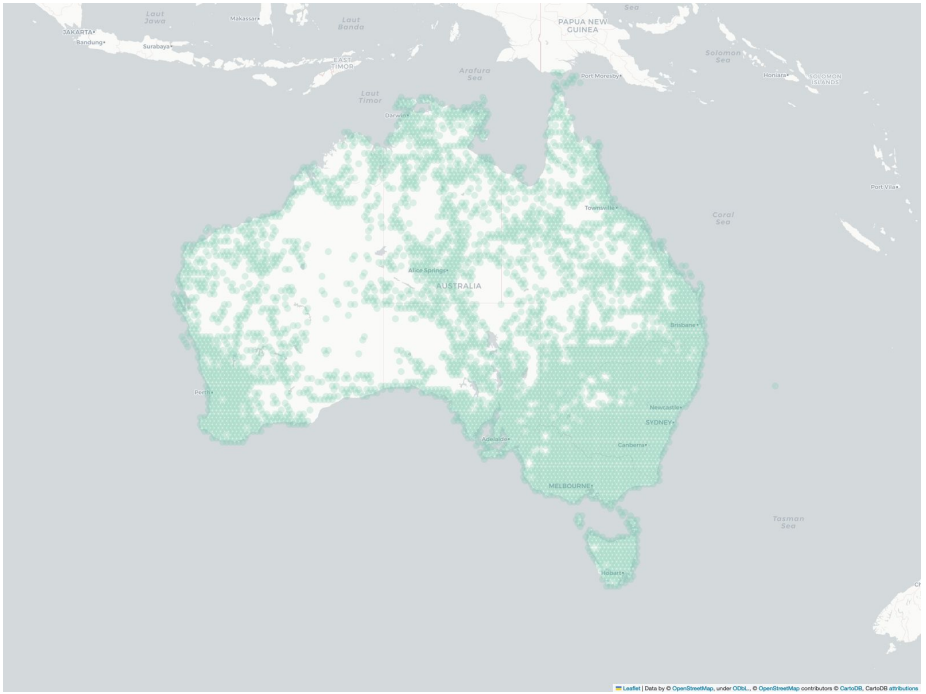
- 34,000 for the US at 10mi radius - so at least 34,000 requests needed
- No result pagination
- Recursive search at smaller radius when there are too many results
- What if we only searched in areas that had buildings?
    - 32% reduction in the USA
    - Now only 23,000 requests



++

Intelligent Coordinate Searching

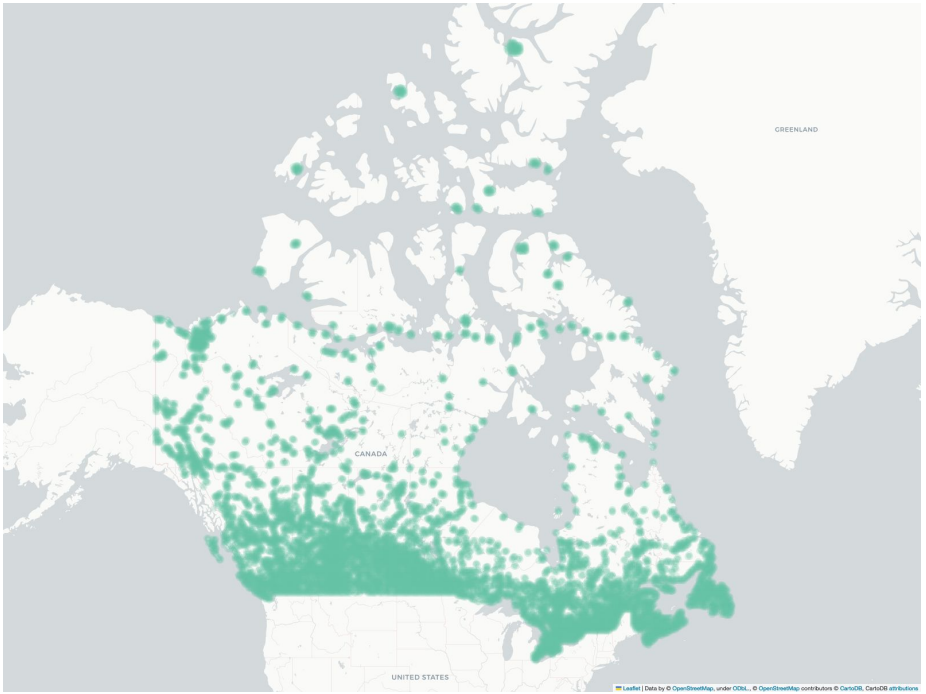
- 34,000 for the US at 10mi radius - so at least 34,000 requests needed
- No result pagination
- Recursive search at smaller radius when there are too many results
- What if we only searched in areas that had buildings?
    - 32% reduction in the USA
    - Now only 23,000 requests
    - 47% reduction in Australia



++

Intelligent Coordinate Searching

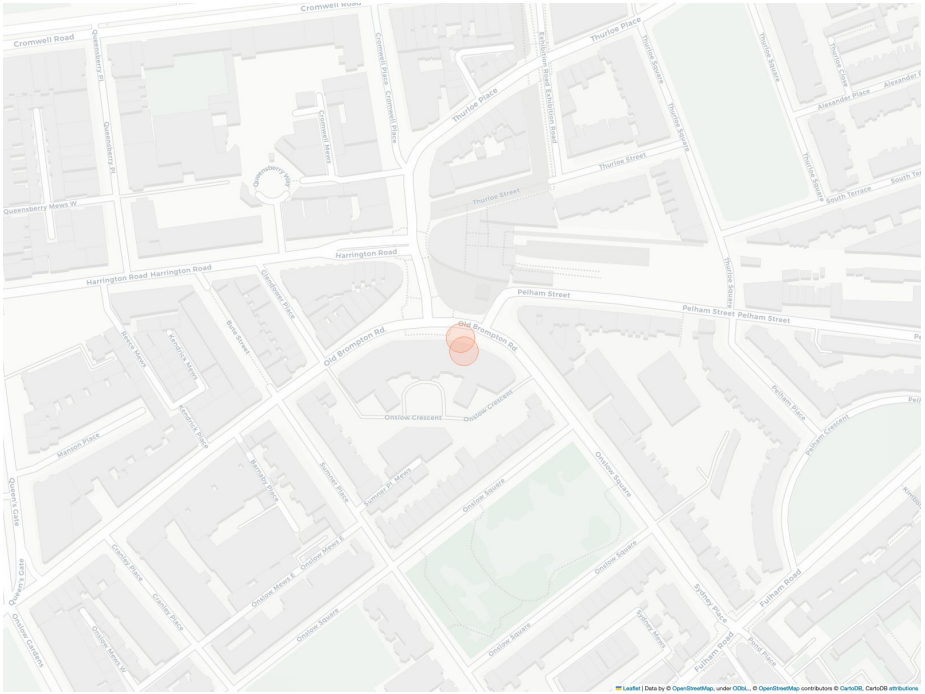
- 34,000 for the US at 10mi radius - so at least 34,000 requests needed
- No result pagination
- Recursive search at smaller radius when there are too many results
- What if we only searched in areas that had buildings?
  - 32% reduction in the USA
  - Now only 23,000 requests
  - 47% reduction in Australia
  - 84% reduction in Canada



++

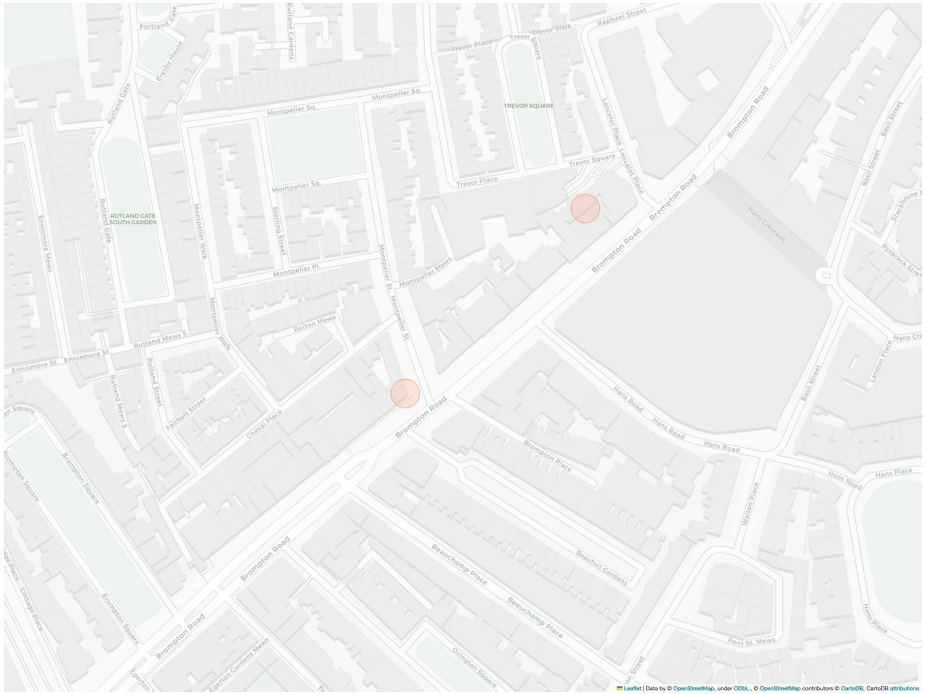
Data Quality Issues

- Duplicates exist even in store finders



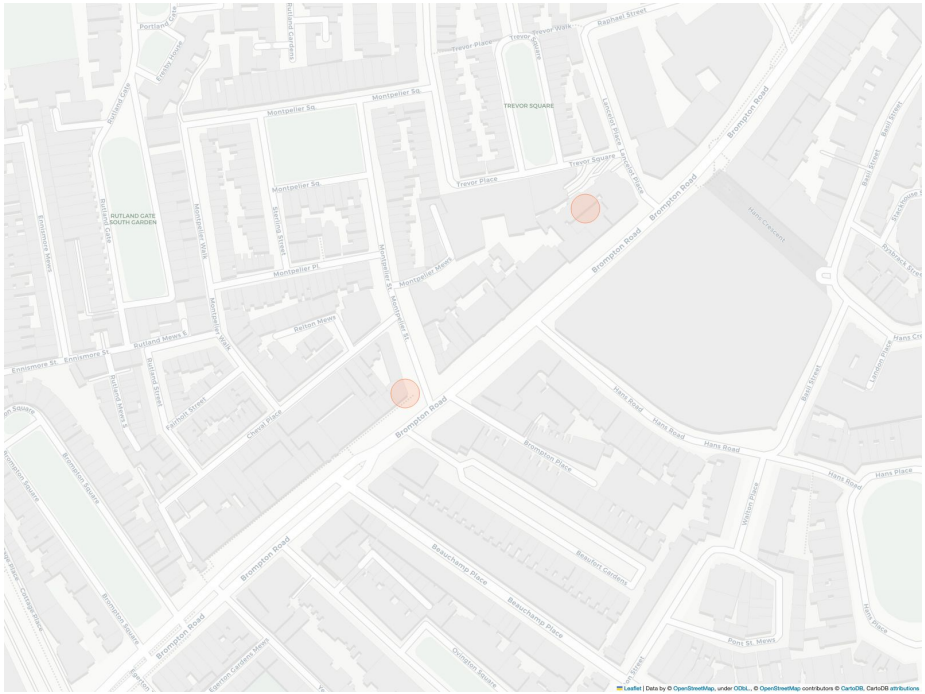
Data Quality Issues

- Duplicates exist even in store finders
  - False positives?



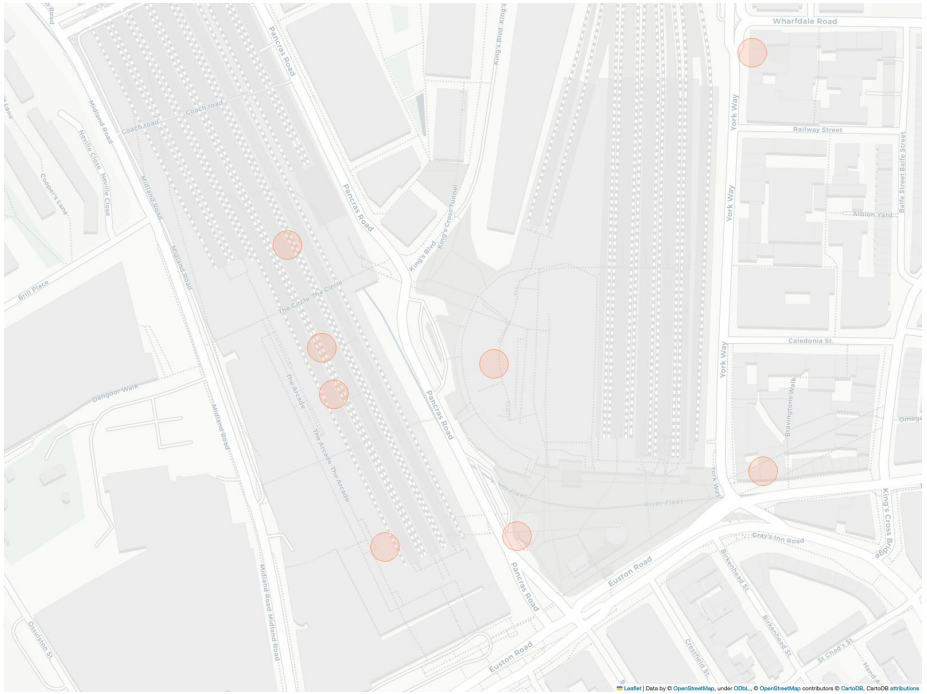
Data Quality Issues

- Duplicates exist even in store finders
  - False positives?
- Brand-provided locations often wrong
  - Metadata, including address, is still correct



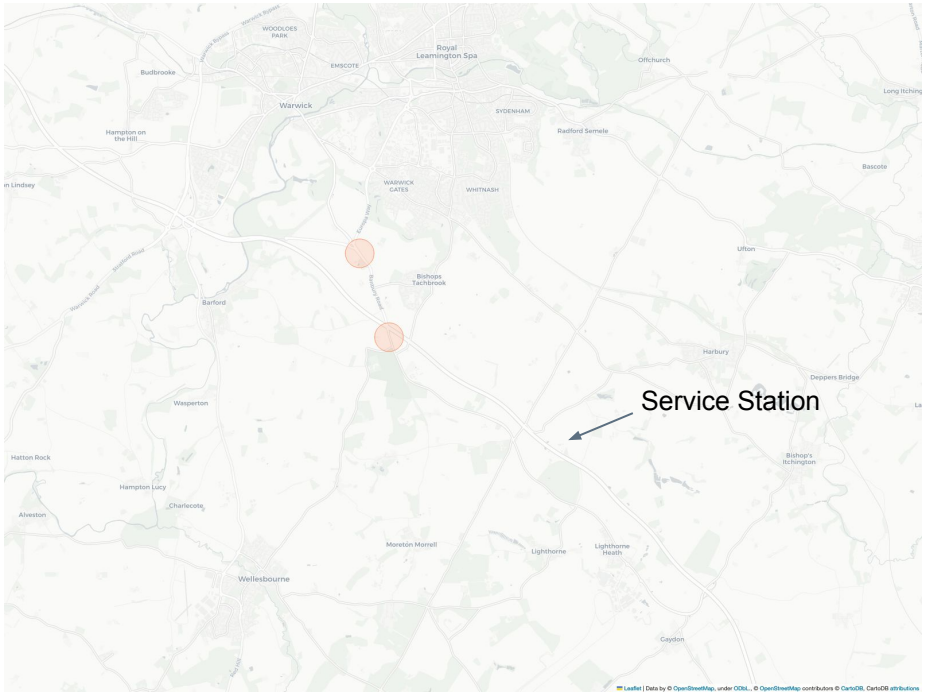
Data Quality Issues

- Duplicates exist even in store finders
- Brand-provided locations often wrong
  - Metadata, including address, is still correct
- Train stations, airports and shopping centers are particular challenge



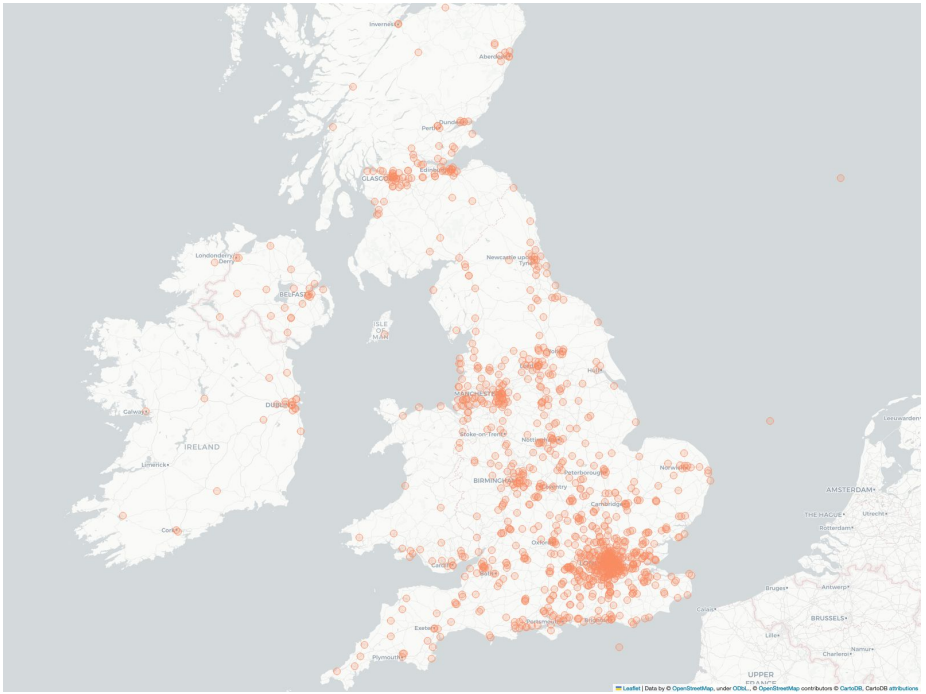
Data Quality Issues

- Duplicates exist even in store finders
- Brand-provided locations often wrong
  - Metadata, including address, is still correct
- Train stations, airports and shopping centers are particular challenge
- Sometimes out by > 1km



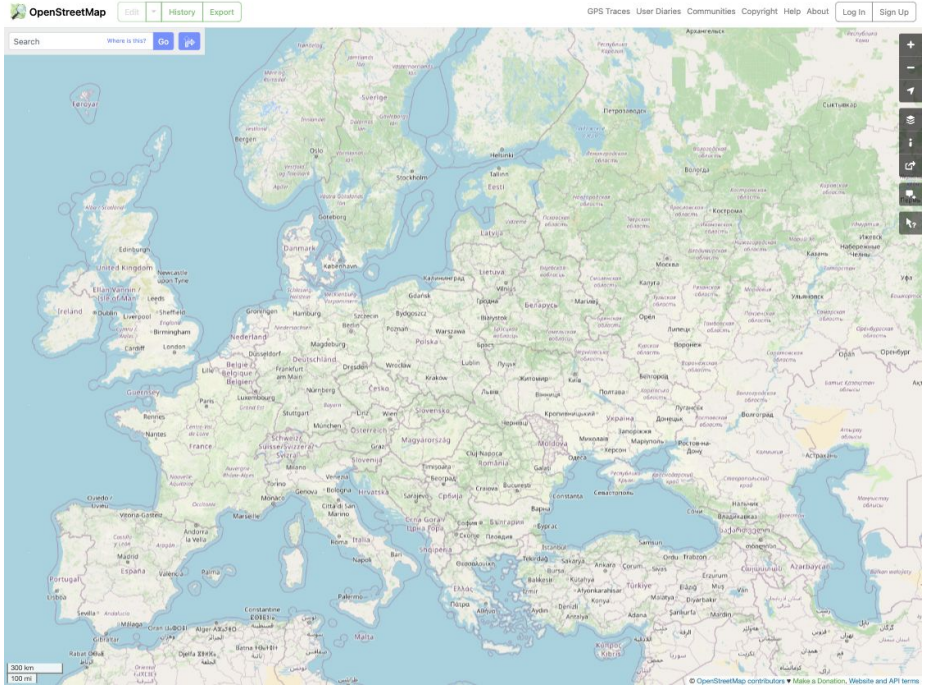
Data Quality Issues

- Duplicates exist even in store finders
- Brand-provided locations often wrong
  - Metadata, including address, is still correct
- Train stations, airports and shopping centers are particular challenge
- Sometimes out by > 1km
- Sometimes ?!?!?



OpenStreetMap

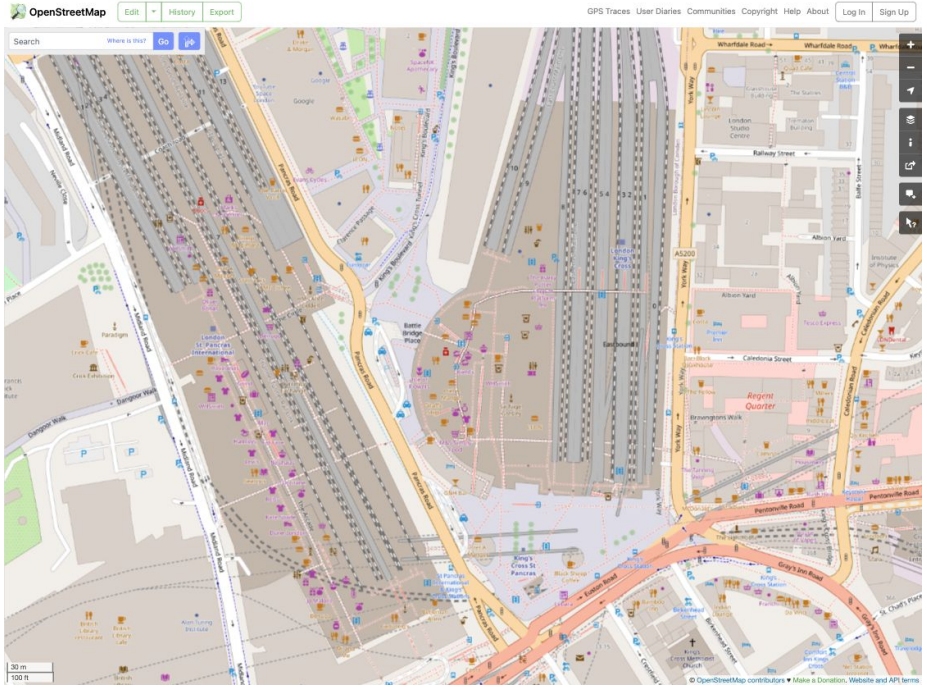
- Crowd-sourced world map
- Released under Open Data Licence



++

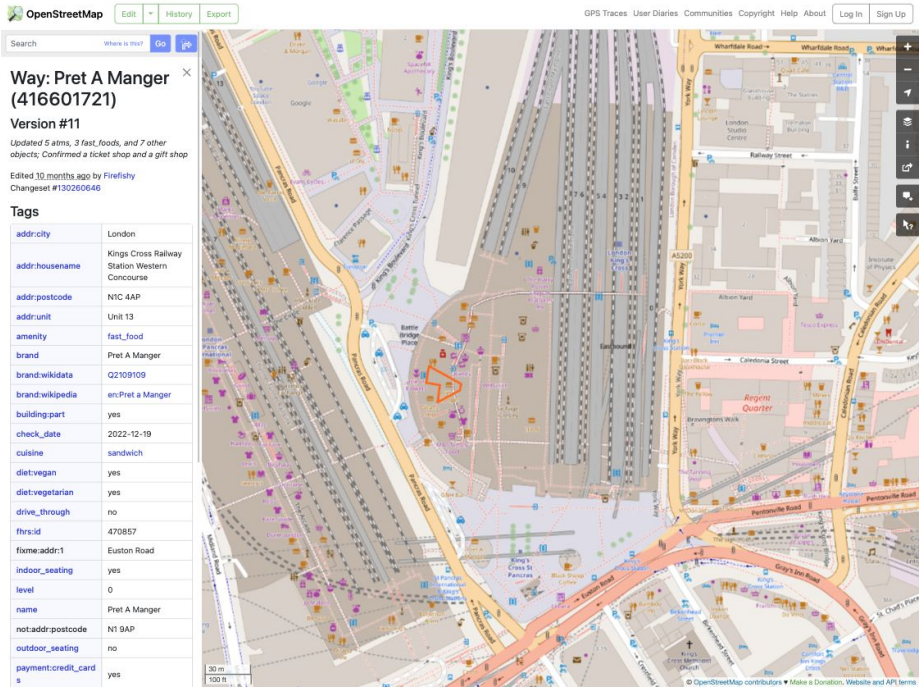
OpenStreetMap

- Crowd-sourced world map
- Released under Open Data Licence
- Community adds POIs in their area



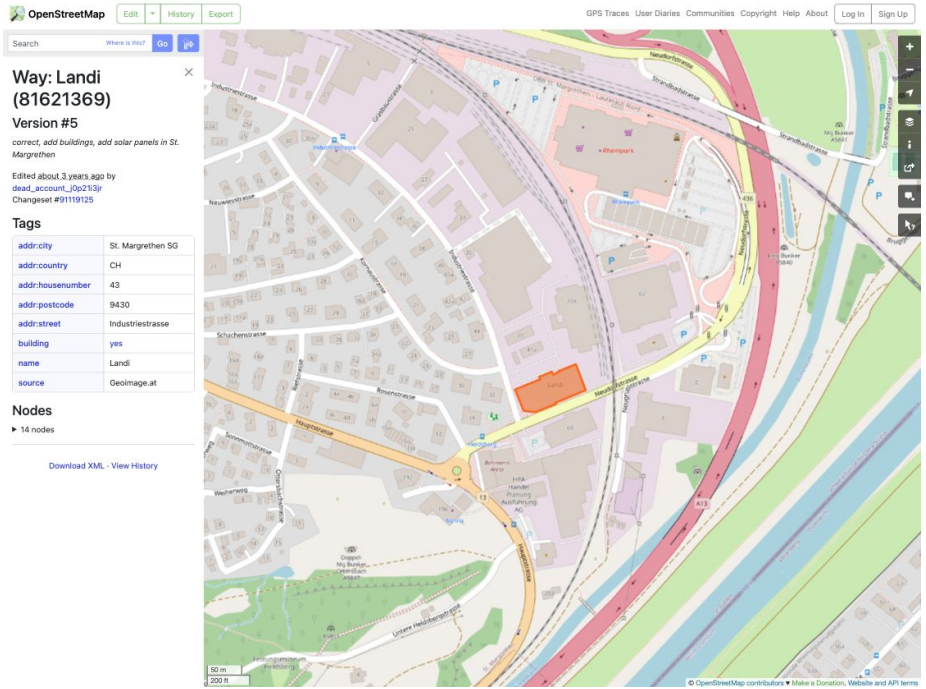
OpenStreetMap

- Crowd-sourced world map
- Released under Open Data Licence
- Footprint polygons of buildings available
- Extremely useful metadata - we can use this to match up with our ATP features
- Has Name Suggestion Index, which auto-suggests metadata based on feature name/brand



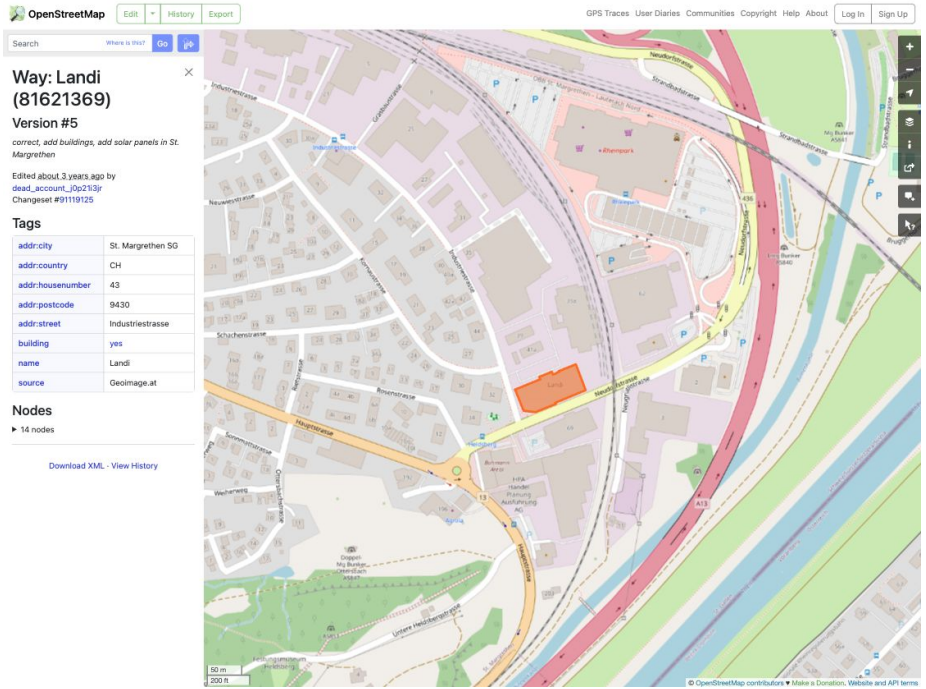
OpenStreetMap

- Crowd-sourced world map
- Released under Open Data Licence
- Footprint polygons of buildings available
- Extremely useful metadata - we can use this to match up with our ATP features
- Has Name Suggestion Index, which auto-suggests metadata based on feature name/brand
- Caveat 1: Metadata not always present
- Caveat 2: Inconsistent tagging



OpenStreetMap

- Crowd-sourced world map
- Released under Open Data Licence
- Footprint polygons of buildings available
- Extremely useful metadata - we can use this to match up with our ATP features
- Has Name Suggestion Index, which auto-suggests metadata based on feature name/brand
- Caveat 1: Metadata not always present
- Caveat 2: Inconsistent tagging
- Caveat 3: Time lag - new branches + + absent / closed branches still present



Let's Match Them Together

- AllThePlaces has:
  - Accurate metadata (if present)
  - Inconsistent lat/lon
  - No polygons
- OpenStreetMap has:
  - Inconsistent metadata
  - … unless editor autocompletion was used, then it's perfect
  - Perfect locations, including building polygons

++

Let's Match Them Together

AllThePlaces

spider	sainsburys
ref	0229
name	Cromwell Road
brand	Sainsbury's
brand_wikidata	Q152096
nsi_id	None
shop	supermarket
amenity	parking
geometry	POINT(-0.18851 51.49527)
country	GB
addr_full	None
located_in	None
located_in_wikidata	None
housenumber	None
street_address	158a Cromwell Road
street	None
city	London
state	None
postcode	SW7 4EJ
website	https://stores.sainsburys.co.uk/0229/cromwell-...
phone	+44 20 7373 8313
opening_hours	Mo-Fr 07:00-23:00; Sa 07:00-22:00; Su 11:00-17:00
image	None
extras	[{'key': 'fhrs:id', 'value': '334461'}, {'key'...

OpenStreetMap

type	way
members	[]
changeset	139821055
timestamp	2023-08-13 12:11:07+00:00
uid	6816132
user	CjMalone
version	14
visible	True
geometry	POLYGON((-0.1889951 51.4951514, -0.1889764 51....
tags	[{'key': 'brand:wikidata', 'value': 'Q152096'}...

Name: 0, dtype: object

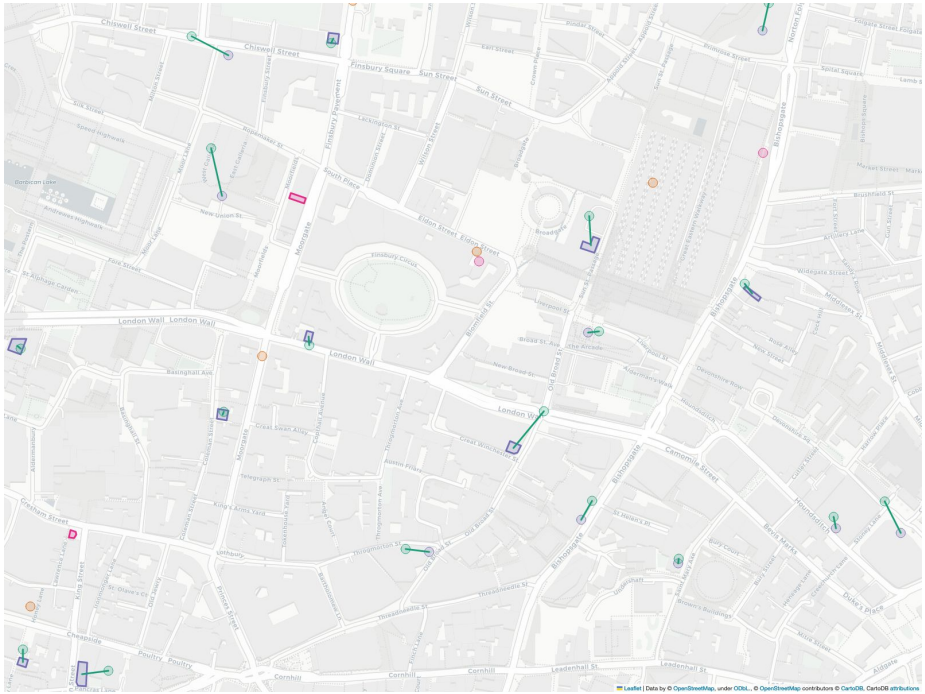
Tags:

name=Sainsbury's
alt_name=Sainsbury's Cromwell Road Superstore
shop=supermarket
building=retail
brand=Sainsbury's
brand:wikipedia=en:Sainsbury's
brand:wikidata=Q152096
addr:city=London
addr:housenumber=158a
addr:street=Cromwell Road
addr:postcode=SW7 4EJ
website=https://stores.sainsburys.co.uk/0229/cromwell-road
opening_hours=Mo 07:00-24:00, Tu-Fr 06:00-24:00, Sa 06:00-22:00; Su 11:00-17:00

Matching - Stage 1

Criteria:

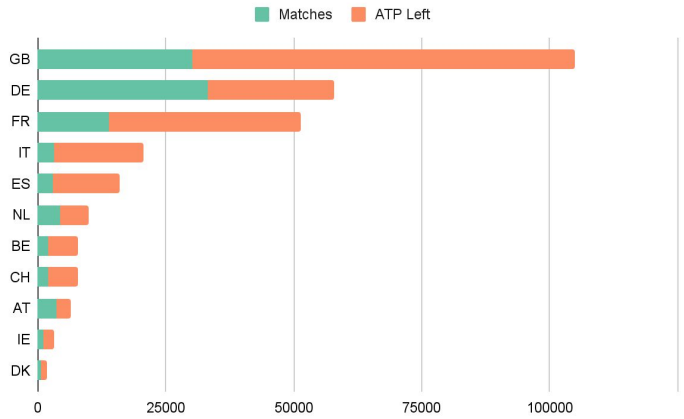
- Within 10m or inside OSM polygon
- `brand_wikidata` matches
  - This indicates that Name Suggestion Index was used on OSM, so metadata is accurate
- `shop` and `amenity` match if present
- Repeat for 20m, 30m, 40m, 50m, 100m for when features are very close to each other



++

Matching - Stage 1 Results

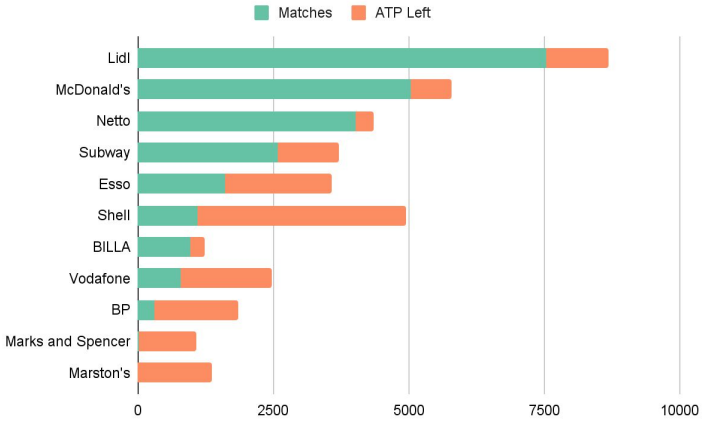
Country	ATP Features	Matches	Match %	Poly Match %	ATP Left	OSM Left
GB	100137	30259	30.22	15.07	74568	18301
DE	58007	33177	57.19	27.12	24830	8582
FR	51374	13948	27.15	11.26	37426	14972
IT	20747	3136	15.12	5.07	17632	3443
ES	15921	2974	18.68	7.96	12957	1826
NL	9811	4429	45.14	2.37	5454	1872
BE	7907	2056	26	11.07	5914	1231
CH	7818	2107	26.95	4.39	5732	795
AT	6431	3682	57.25	25.69	2749	814
IE	3171	1203	37.94	17.98	1969	595
DK	1833	769	41.95	19.97	1064	360



++

Matching - Stage 1 Results

Brand	ATP Features	Matches	Match %	Poly %	ATP Left
Lidl	8687	7534	86.73	59.17	1153
McDonald's	5798	5046	87.03	52.24	752
Netto	4352	4016	92.28	59.95	336
Subway	3712	2591	69.8	14.09	1121
Esso	3569	1618	45.33	20.12	1951
Shell	4951	1103	22.28	10.44	3848
BILLA	1236	980	79.29	38.35	256
Vodafone	2466	785	31.83	1.74	1681
BP	1857	301	16.21	5.65	1556
Marks and Spencer	1078	12	1.11	0.37	1066
Marston's	1364	4	0.29	0.15	1360



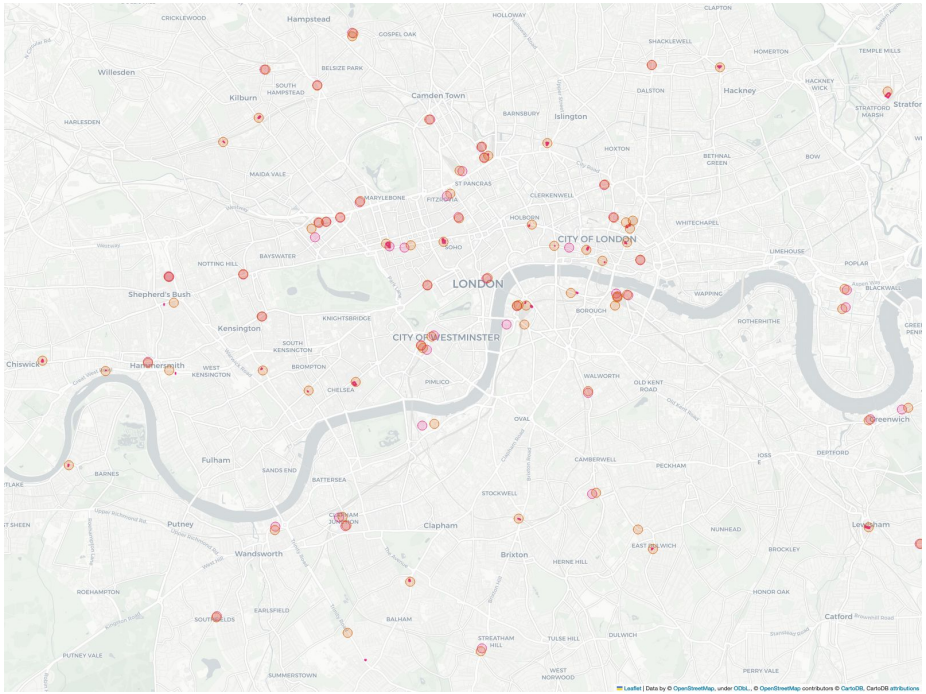
++

Problem brand: Marks and Spencer

ATP and OSM disagree on how to tag

Brand	Count
Marks and Spencer	682
M&S Simply Food	245
M&S Outlet	29
M&S Foodhall	99

Brand	Count
Marks and Spencer	3
Marks & Spencer Food	1
Marks & Spencer	334
M&S Simply Food	420
M&S Outlet	22
M&S Home	1
M&S Foodhall	196
M&S Food To Go	1
M&S Food	3
null	7



++

Solution: Stage 2 - Fuzzy Matching with LLMs

- Use LLM for our “best guesses”
- Google Vertex AI
- 300m range
- Pass metadata including:
  - Store name
  - Brand
  - Address
  - Phone
  - Website
  - Feature type (`shop/amenity`)
- Try the closest three ATP features for every OSM feature
- Allows flexibility, e.g.

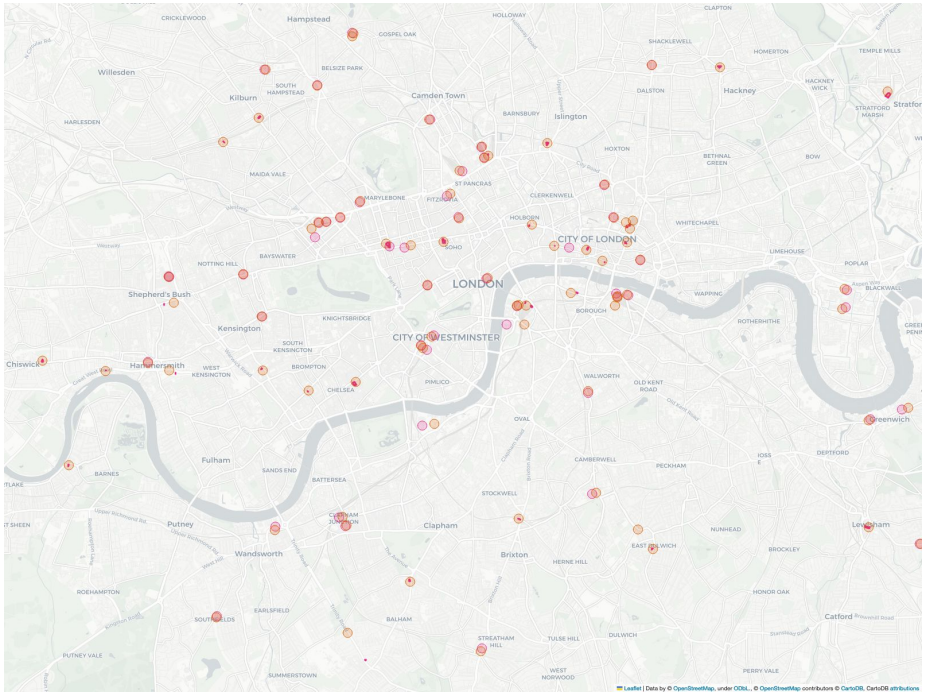
`shop=supermarket` **vs** `shop=convenience`

++

```
Your task is to find matches between table rows that represent
points of interest obtained from two different sources.
You are given a row that should be matched and up to 3 merging
candidates, enumerated with an integer index provided in
parenthesis. You get the distance between the geometry centroid
and values in JSON format. Fields with null value have been
omitted for brevity.

Row to be matched:
    values:{"name":"M&S Foodhall","brand":"M&S
Foodhall","shop":"supermarket"}

Match candidates:
(1): distance: 11m
    values: {"name":"LONDON WATERLOO RAIL SIMPLY
FOOD","brand":"M&S Simply Food","city":"London","postcode":"SE1
7LY"}
(2): distance: 46m
    values: {"name":"SOUTHBANK PLACE","brand":"Marks and
Spencer","city":"London","postcode":"SE1 7ND"}
(3): distance: 199m
    values: {"name":"Waterloo Station South","brand":"Marks and
Spencer","city":"London","postcode":"SE1 8SW"}

Please respond with the integer index if one of the candidates is
a plausible match, otherwise return 0.
```

Solved brand: Marks and Spencer

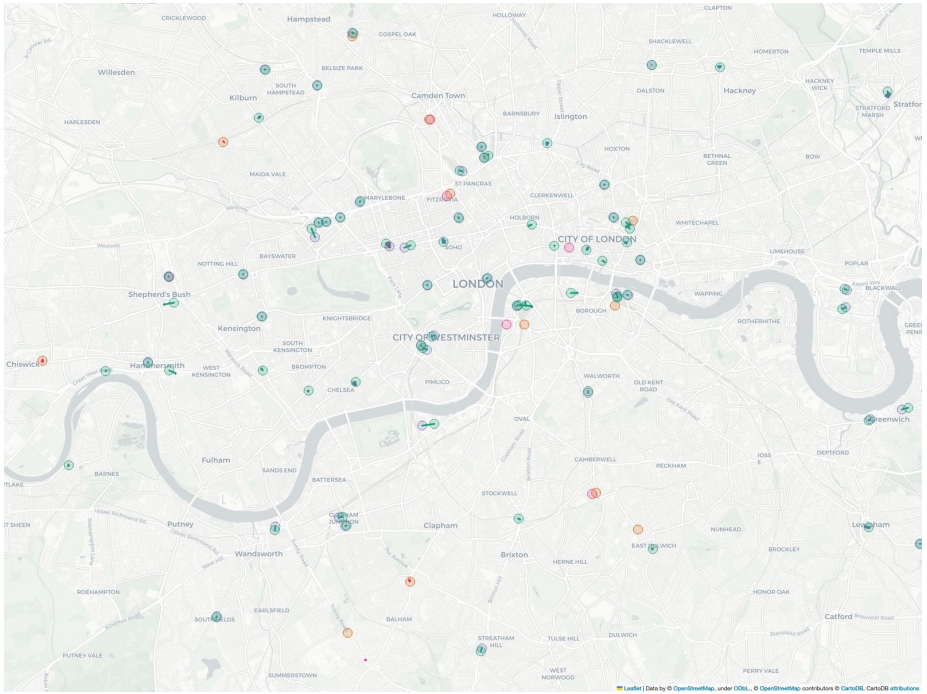
- VertexAI correctly assigns ATP and OSM features based on distance and metadata
- Seems obvious for Marks and Spencers… imagine it for Starbucks in NYC



++

Solved brand: Marks and Spencer

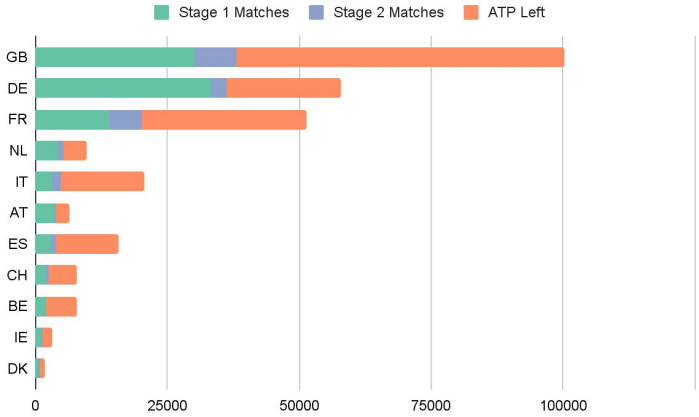
- VertexAI correctly assigns ATP and OSM features based on distance and metadata
- Seems obvious for Marks and Spencers… imagine it for Starbucks in NYC



++

Matching - Stage 2 Results

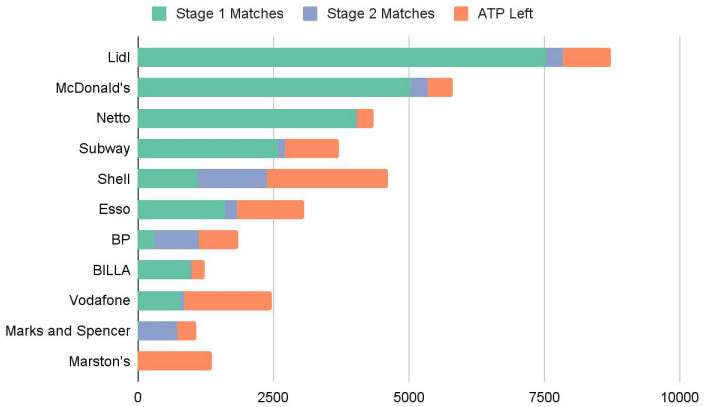
Country	ATP Features	Stage 1 Matches	Stage 2 Matches	Stage 2 Match %	Overall Match %	ATP Left
GB	100137	30259	7859	7.85	38.07	62019
DE	58007	33177	3142	5.42	62.61	21688
FR	51374	13948	6255	12.18	39.33	31171
NL	9811	4429	899	9.16	54.31	4483
IT	20747	3136	1651	7.96	23.07	15960
AT	6431	3682	373	5.80	63.05	2376
ES	15921	2974	1007	6.32	25.00	11940
CH	7818	2107	533	6.82	33.77	5178
BE	7907	2056	87	1.10	27.10	5764
IE	3171	1203	240	7.57	45.51	1728



++

Matching - Stage 2 Results

Country	ATP Features	Stage 1 Matches	Stage 2 Matches	Stage Match %	Match %
Lidl	8687	7534	316	3.64	90.36
McDonald's	5798	5046	308	5.31	92.34
Netto	4352	4016	32	0.74	93.01
Subway	3712	2591	124	3.34	73.14
Shell	4951	1103	1277	**25.79**	48.07
Esso	3569	1618	212	5.94	51.27
BP	1857	301	836	**45.02**	61.23
BILLA	1236	980	20	1.62	80.91
Vodafone	2466	785	75	3.04	34.87
Marks and Spencer	1078	12	721	**66.88**	68.00



Problem brand: Marston's Pubs

- Pubs lead with their name, not brand
- Mappers do not attach brand when contributing to OpenStreetMap

Way: The Smugglers Cove (669720795)

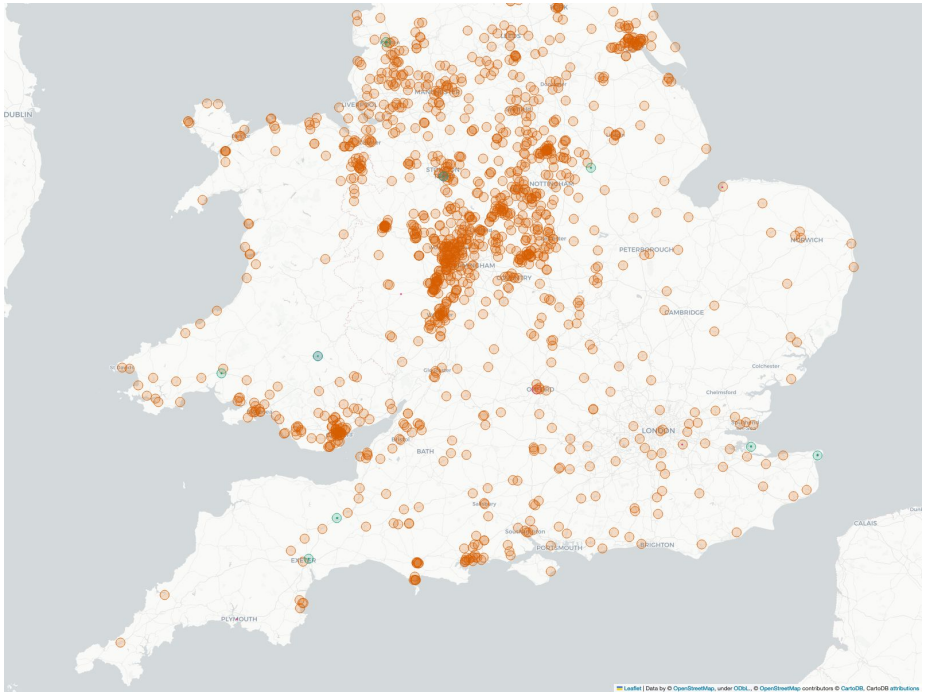
Version #4

Improve Brook Park West alignment using updated Bing imagery

Edited about 3 years ago by EdLoach
Changeset #89566807

Tags

addr:city	Clacton-on-Sea
addr:housename	The Smugglers Cove
addr:postcode	CO16 9GA
addr:street	Hartley Brook Road
amenity	pub
building	pub
name	The Smugglers Cove



Solution: Query OpenStreetMap by name

- `brand_wikidata` is no longer useful - the OSM feature isn't tagged with it

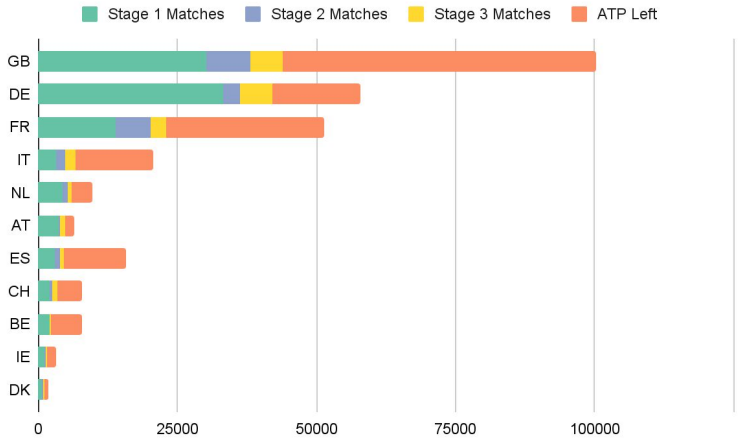
New criteria:

- Search full OSM features
- Within 200m
- ATP name or brand matches OSM name
- `shop` and `amenity` match if present

++

Matching - Stage 3 Results

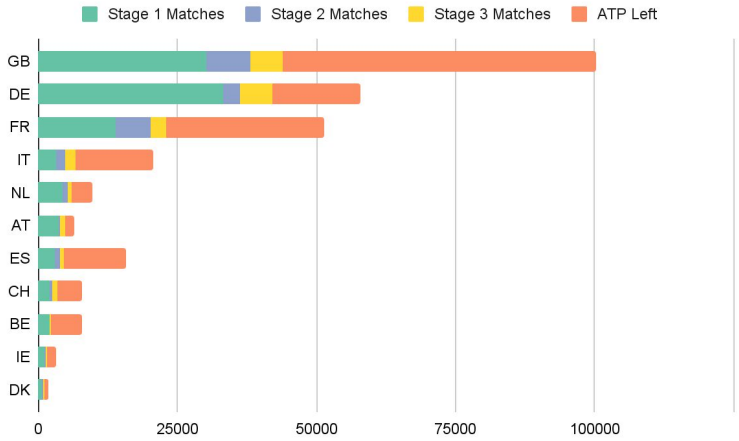
Country	ATP Features	Stage 1 Matches	Stage 2 Matches	Stage 3 Matches	Stage 3 Match %	Overall Match %	ATP Left
GB	100137	30259	7859	5764	5.76	43.82	56255
DE	58007	33177	3142	5867	10.11	72.73	15821
FR	51374	13948	6255	2760	5.37	44.70	28411
IT	20747	3136	1651	2035	9.81	32.88	13925
NL	9811	4429	899	792	8.07	62.38	3691
AT	6431	3682	373	709	11.02	74.08	1667
ES	15921	2974	1007	736	4.62	29.63	11204
CH	7818	2107	533	884	11.31	45.08	4294
BE	7907	2056	87	110	1.39	28.49	5654
IE	3171	1203	240	167	5.27	50.77	1561



++

Matching - Stage 3 Results

Country	ATP Features	Stage 3 Matches	Stage Match %	Match %	ATP Left
Lidl	8687	469	5.40	95.76	368
McDonald's	5798	38	0.66	93.00	406
Netto	4352	103	2.37	95.38	201
Shell	4951	1140	**23.03**	71.10	1431
Subway	3712	48	1.29	74.43	949
Esso	3569	335	9.39	60.66	1404
BP	1857	161	8.67	69.90	559
BILLA	1236	193	**15.61**	96.52	43
Vodafone	2466	271	**10.99**	45.86	1335
Marks and Spencer	1078	30	2.78	70.78	315



++

Still a problem brand: Marston's Pubs

- Pubs lead with name, not their brand
- Mappers do not attach brand when contributing to OpenStreetMap

Way: The Smugglers Cove (669720795)

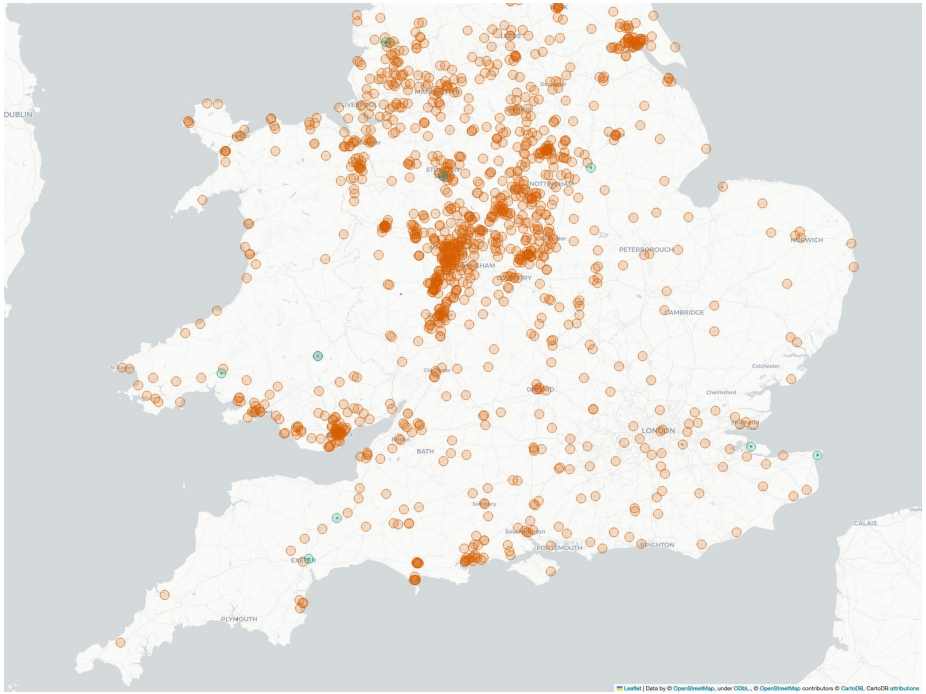
Version #4

Improve Brook Park West alignment using updated Bing imagery

Edited about 3 years ago by EdLoach
Changeset #89566807

Tags

addr:city	Clacton-on-Sea
addr:housename	The Smugglers Cove
addr:postcode	CO16 9GA
addr:street	Hartley Brook Road
amenity	pub
building	pub
name	The Smugglers Cove



Still a problem brand: Marston's Pubs

- Some improvement
- … but we need more flexibility

Way: The Smugglers Cove (669720795)

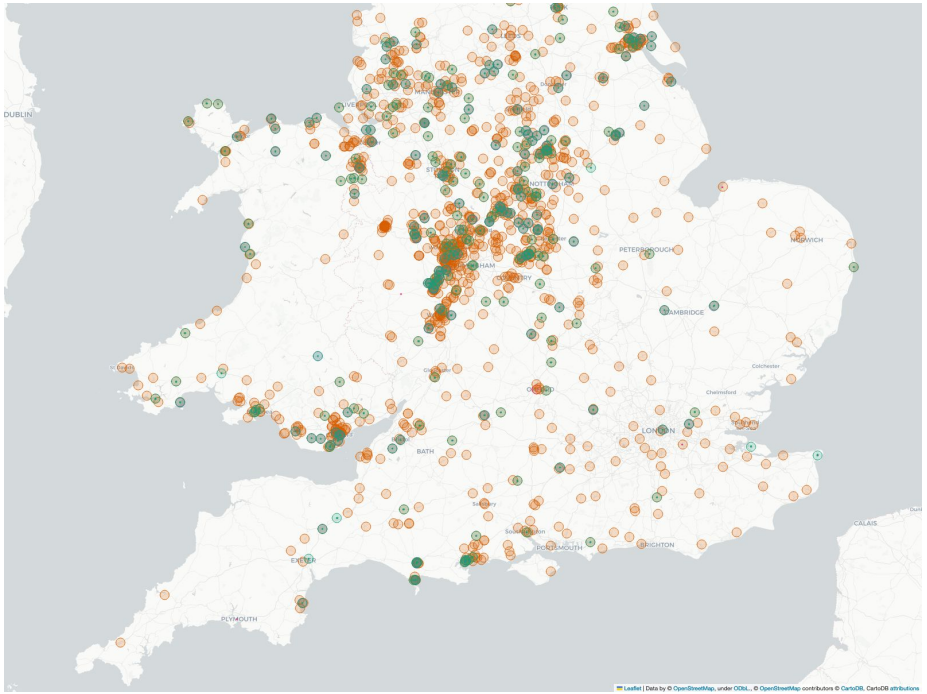
Version #4

Improve Brook Park West alignment using updated Bing imagery

Edited about 3 years ago by EdLoach
Changeset #89566807

Tags

addr:city	Clacton-on-Sea
addr:housename	The Smugglers Cove
addr:postcode	CO16 9GA
addr:street	Hartley Brook Road
amenity	pub
building	pub
name	The Smugglers Cove



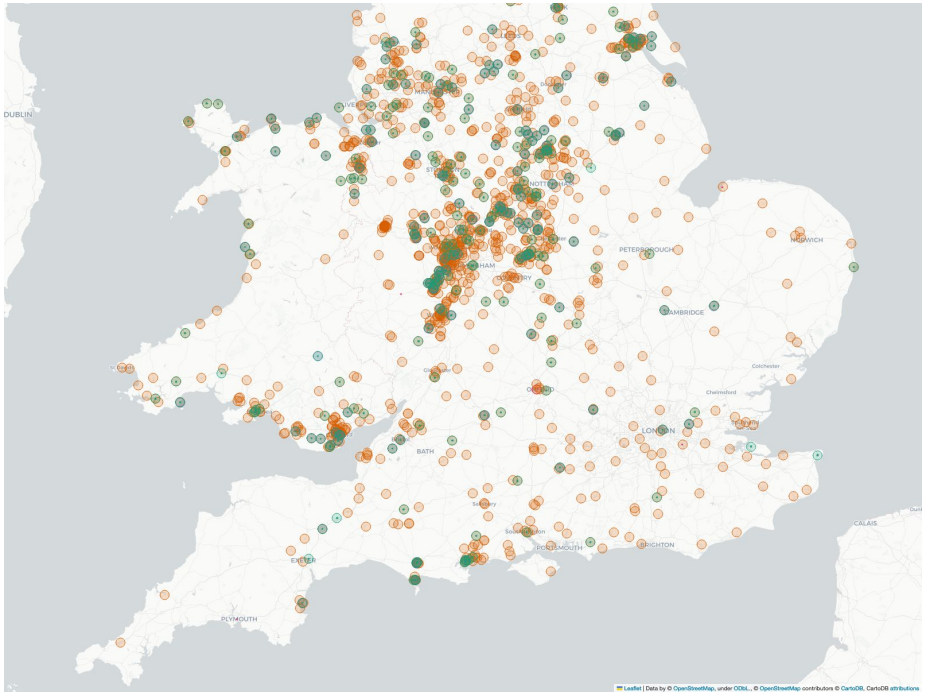
Solution: Stage 4 - VertexAI and OSM combined

- Take advantage of the LLM's strengths in text parsing
- Feed in every OSM POI within 200m
- Let the LLM decide if there is a relevant match

```
.......
Row to be matched:
    values:{"name":"Wilmslow
Tavern","brand":"Marston's","amenity":"pub","city":"Wilmslow","sta
te":"Cheshire","postcode":"SK9 2HA"}

Match candidates:
(1):
    values: {"name":"Village Centre Dry
Cleaners","shop":"dry_cleaning"}
(2):
    values: {"name":"PodPoint Charging
Station","amenity":"charging_station"}
(3):
    values: {"name":"Tanning & Beauty","shop":"beauty"}
(4):
    values: {"name":"The Wilmslow Tavern","amenity":"pub"}
(5):
    values: {"name":"Marsha Lloyd","shop":"hairdresser"}
(6):
    values: {"name":"Lidl","brand":"Lidl","shop":"supermarket"}
(7:
    values: {"name":"Colshaw Farm"}
(8):
    values: {"name":"Lloyds Pharmacy","brand":"Lloyds
Pharmacy","amenity":"pharmacy"}
```

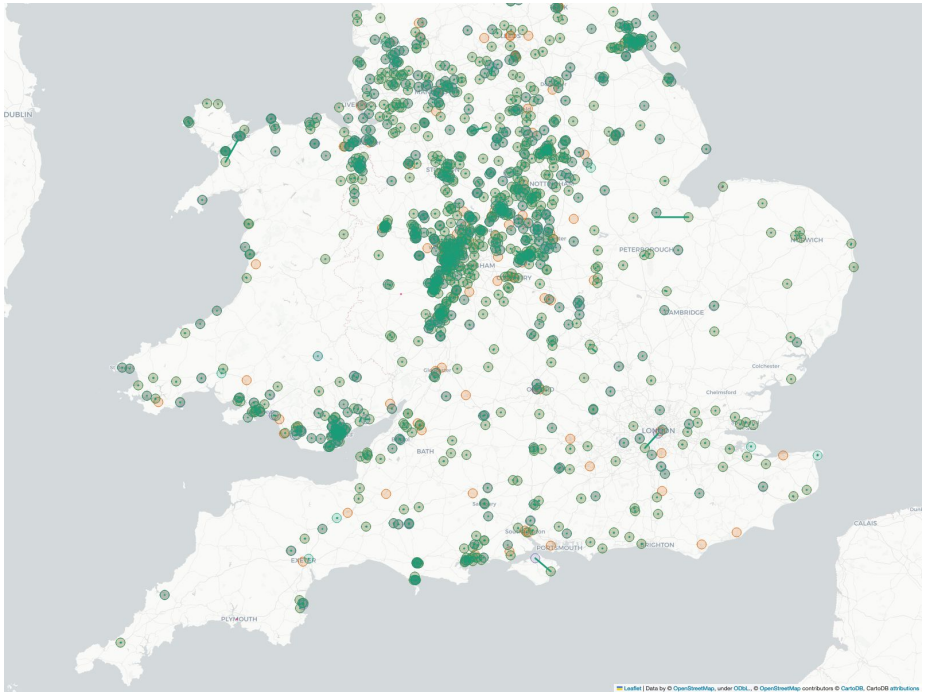
Fixed? Marston's Pubs

- We've found the majority of the pubs!



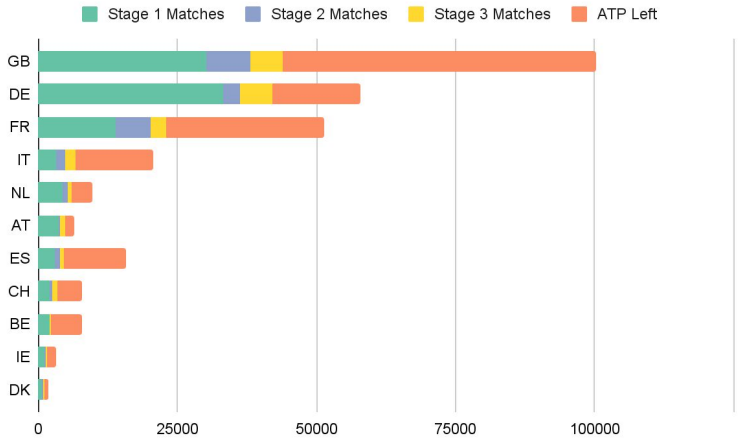
Fixed? Marston's Pubs

- We've found the majority of the pubs!
- We're taking more risks with mismatches now
- Perfection is never going to be achievable, but we'd rather have POIs missing than in the wrong place



Matching - Stage 4 Results

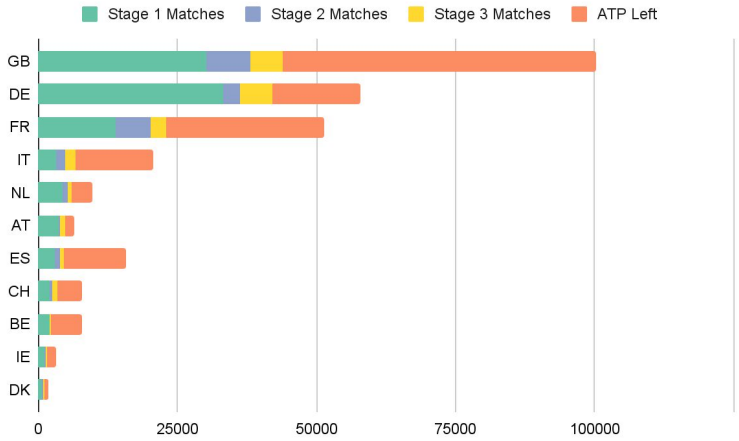
Country	ATP Features	Stage 1 Matches	Stage 2 Matches	Stage 3 Matches	Stage 3 Match %	Overall Match %	ATP Left
GB	100137	30259	7859	5764	5.76	43.82	56255
DE	58007	33177	3142	5867	10.11	72.73	15821
FR	51374	13948	6255	2760	5.37	44.70	28411
IT	20747	3136	1651	2035	9.81	32.88	13925
NL	9811	4429	899	792	8.07	62.38	3691
AT	6431	3682	373	709	11.02	74.08	1667
ES	15921	2974	1007	736	4.62	29.63	11204
CH	7818	2107	533	884	11.31	45.08	4294
BE	7907	2056	87	110	1.39	28.49	5654
IE	3171	1203	240	167	5.27	50.77	1561



++

Matching - Stage 4 Results

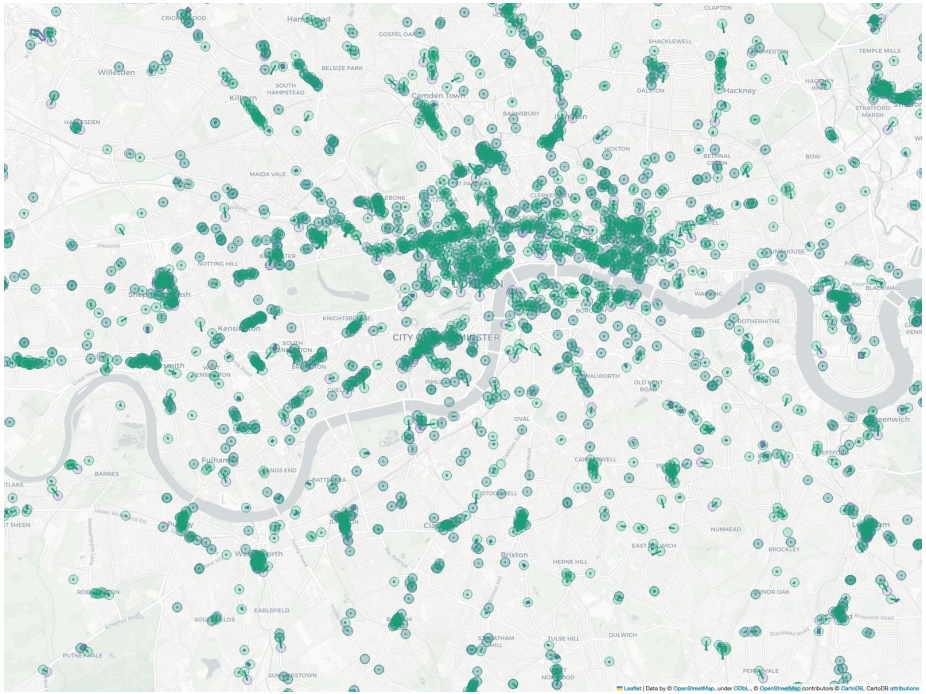
Country	ATP Features	Stage 3 Matches	Stage Match %	Match %	ATP Left
Lidl	8687	469	5.40	95.76	368
McDonald's	5798	38	0.66	93.00	406
Netto	4352	103	2.37	95.38	201
Shell	4951	1140	**23.03**	71.10	1431
Subway	3712	48	1.29	74.43	949
Esso	3569	335	9.39	60.66	1404
BP	1857	161	8.67	69.90	559
BILLA	1236	193	**15.61**	96.52	43
Vodafone	2466	271	**10.99**	45.86	1335
Marks and Spencer	1078	30	2.78	70.78	315



++

Next Steps

- Contribute to OpenStreetMap!
  - Submit our edit suggestions to OSM quest apps
  - Suggest new branches
  - Identify closed branches
- Align ATP more closely with OSM
- Write many, many more spiders
- Give the LLM examples of good matches
- Use ATP and OSM history to narrow down branch open/close dates
- Relax and enjoy the view…



++

The End Product - Thanks for Watching!

++