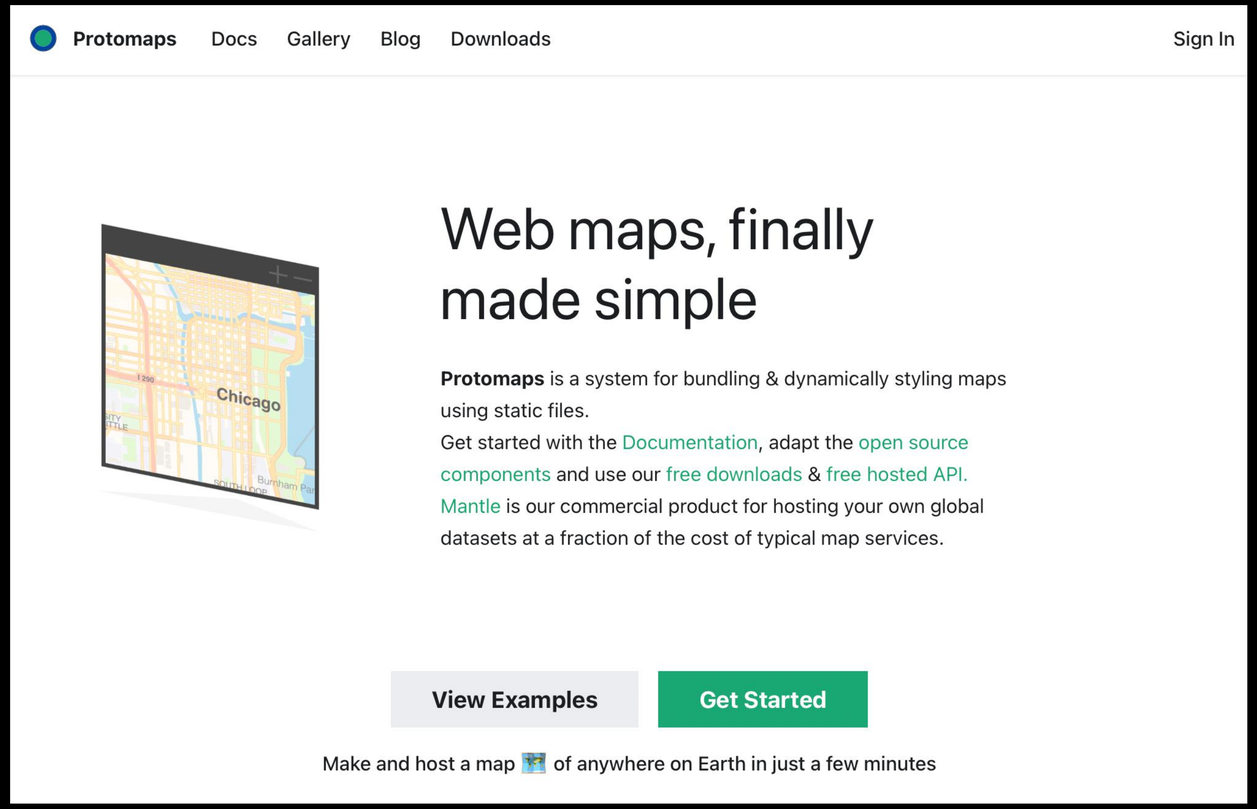
Innovating on Derivative OSM Datasets

Brandon Liu / Protomaps LLC / State of the Map 2022 Firenze

Background

Protomaps

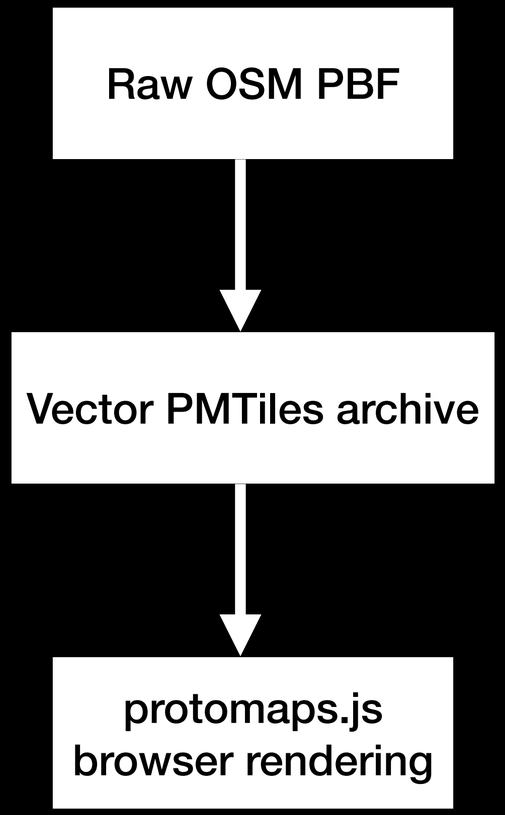
End-to-end toolchain for custom self-hosted vector basemaps



Background

- **Core values:** simplicity + easy to self-host
  - Vector basemap tiles, No PostGIS, no tileservers, S3 only
- Several independent **open-source subprojects on github.com/protomaps**
  - Rendering = protomaps.js (canvas + leaflet + typescript)
  - OSM database = OSM Express
  - Cloud-optimized tile archives = PMTiles

Protomaps Toolchain



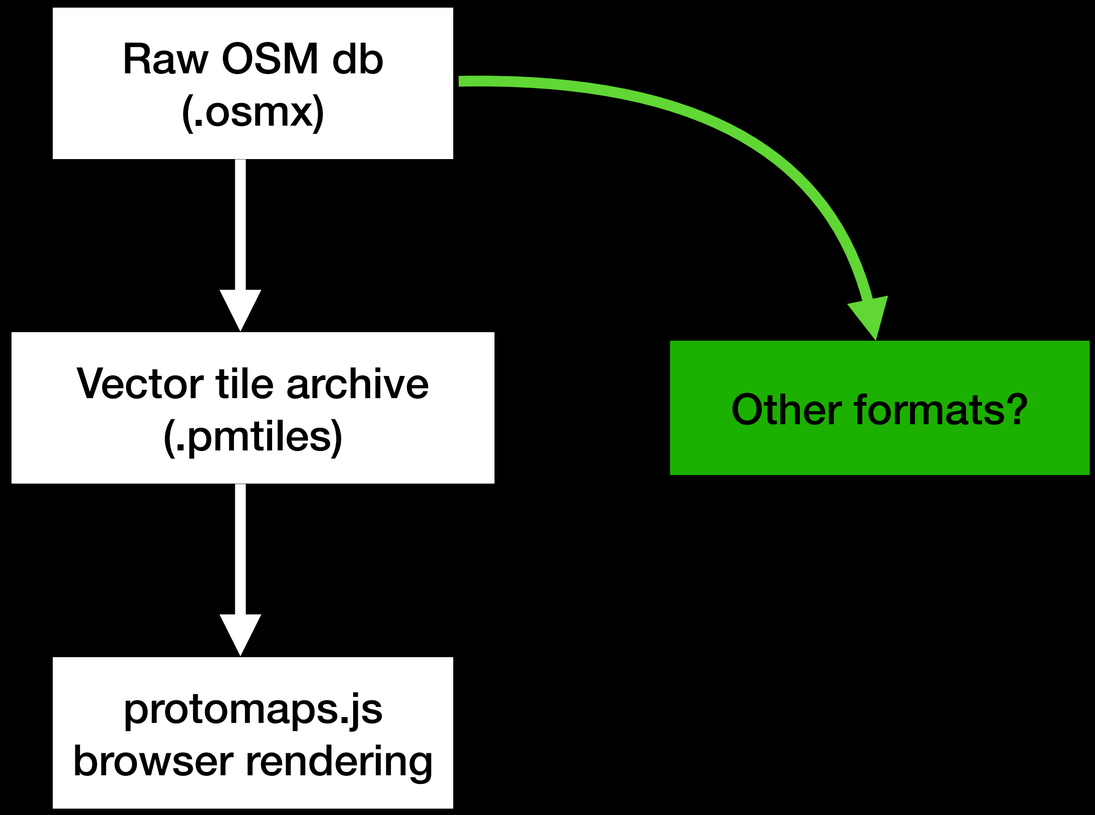
More than tiles

- Among OSM data products, tiles attract the most interest
- **But,** tiles are a specific display-oriented technique; less suited to other tasks

More than tiles

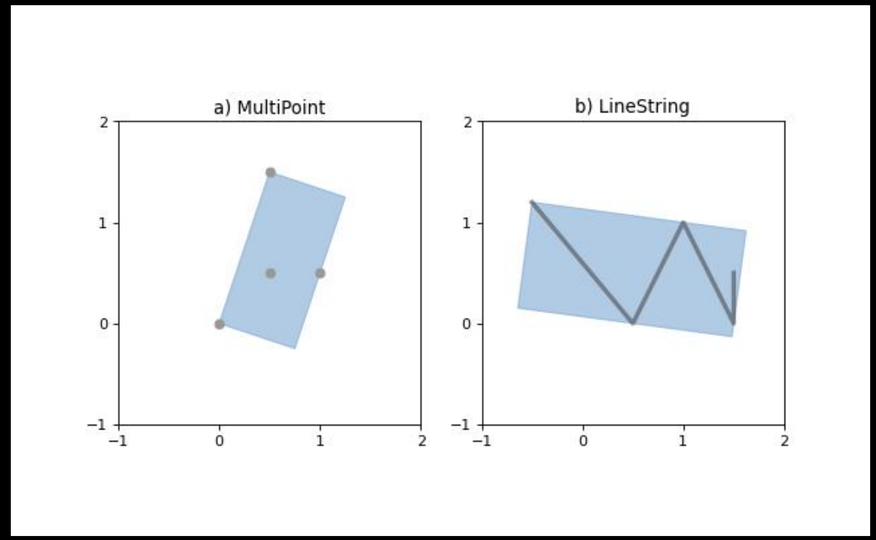
- Datasets for use in QGIS, ArcGIS
- Analytical tasks need complete unsimplified geometries
- Input for making print and non-web maps

Protomaps Toolchain



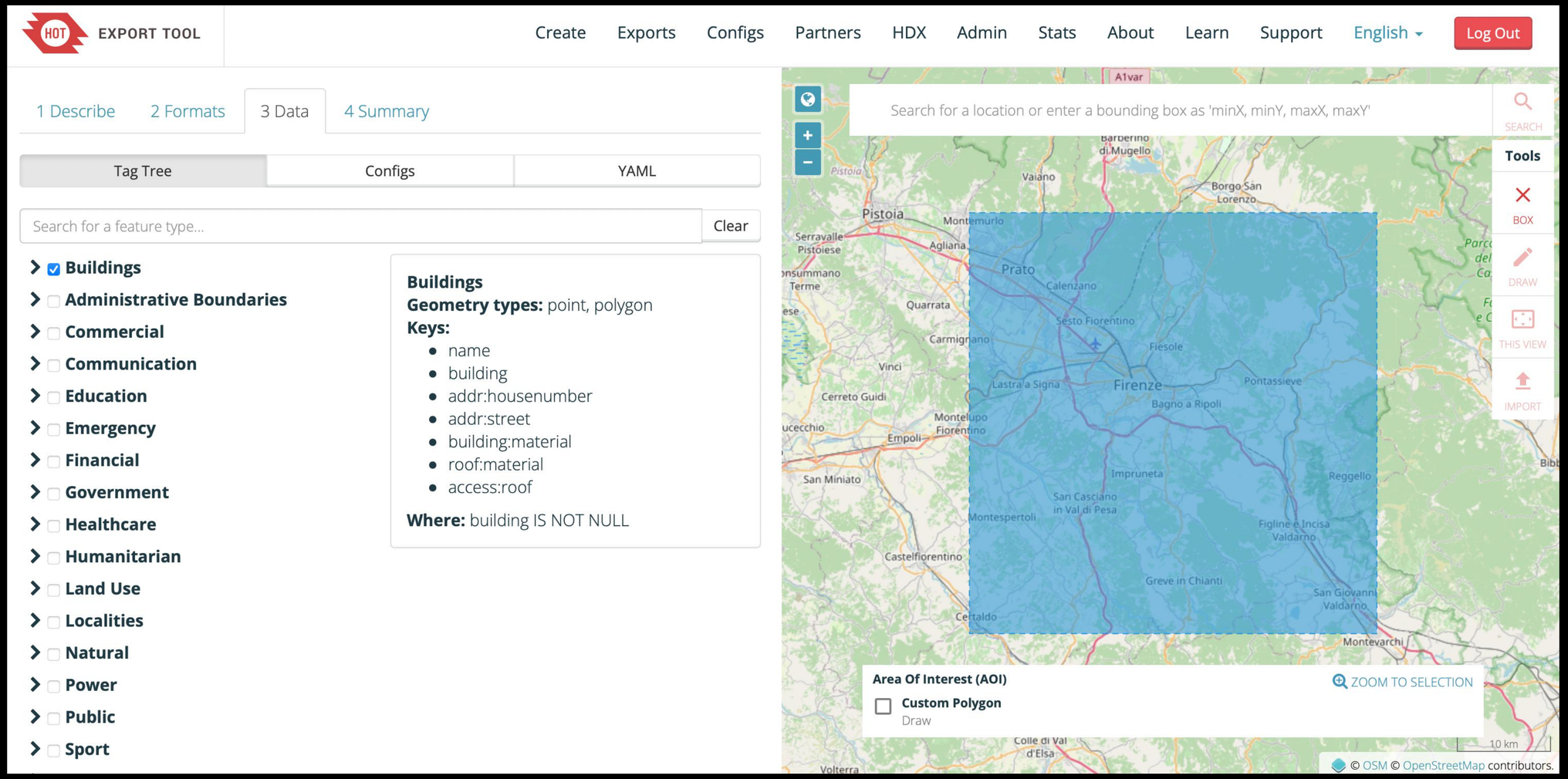
OpenStreetMap as GIS Data?

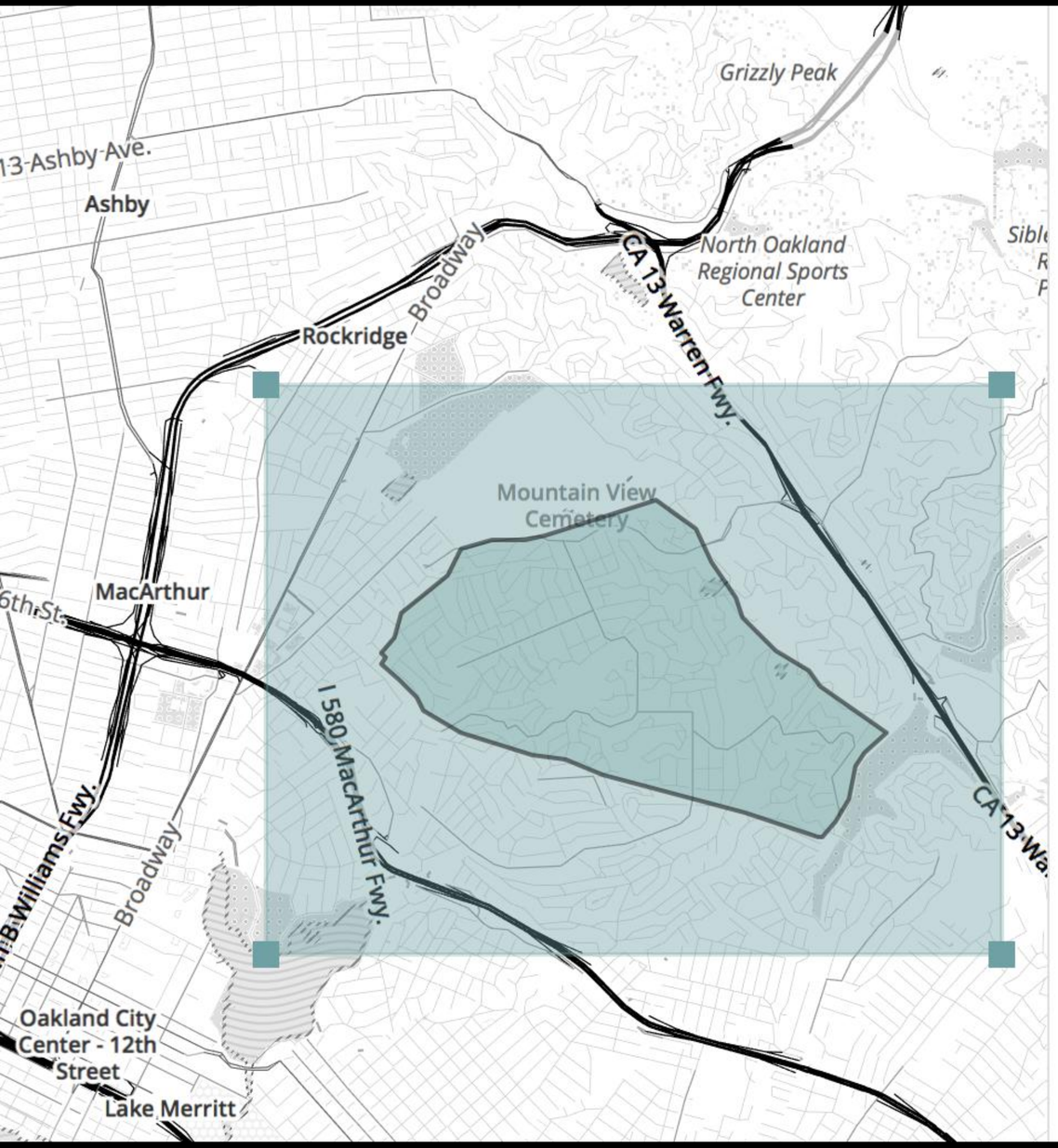
- For an audience of GIS users in general (a large group)
- Not for an audience of OSM experts (a small group)
- Deliver data in the format GIS users are used to
  - **OGC Simple Features**
  - Pre-selected areas vs on-demand



The Shapely Manual ❤️

HOTOSM Export Tool





metro extracts

Each week, Metro Extracts automatically extracts the latest OpenStreetMap data into manageable, metro-area files in a variety of formats for you to use.

The red boxes represent existing extracts, which you can download right now. You can also request a new extract (shown as a blue box) which will take 30-60 minutes to create.

Documentation | Tutorial | File Format Guide | Your Extracts

SEARCH FOR A CITY OR REGION

Piedmont, CA, USA

SEARCH

Your extract is included inside of a larger area:

ENCOMPASSING METROS

San Francisco Bay

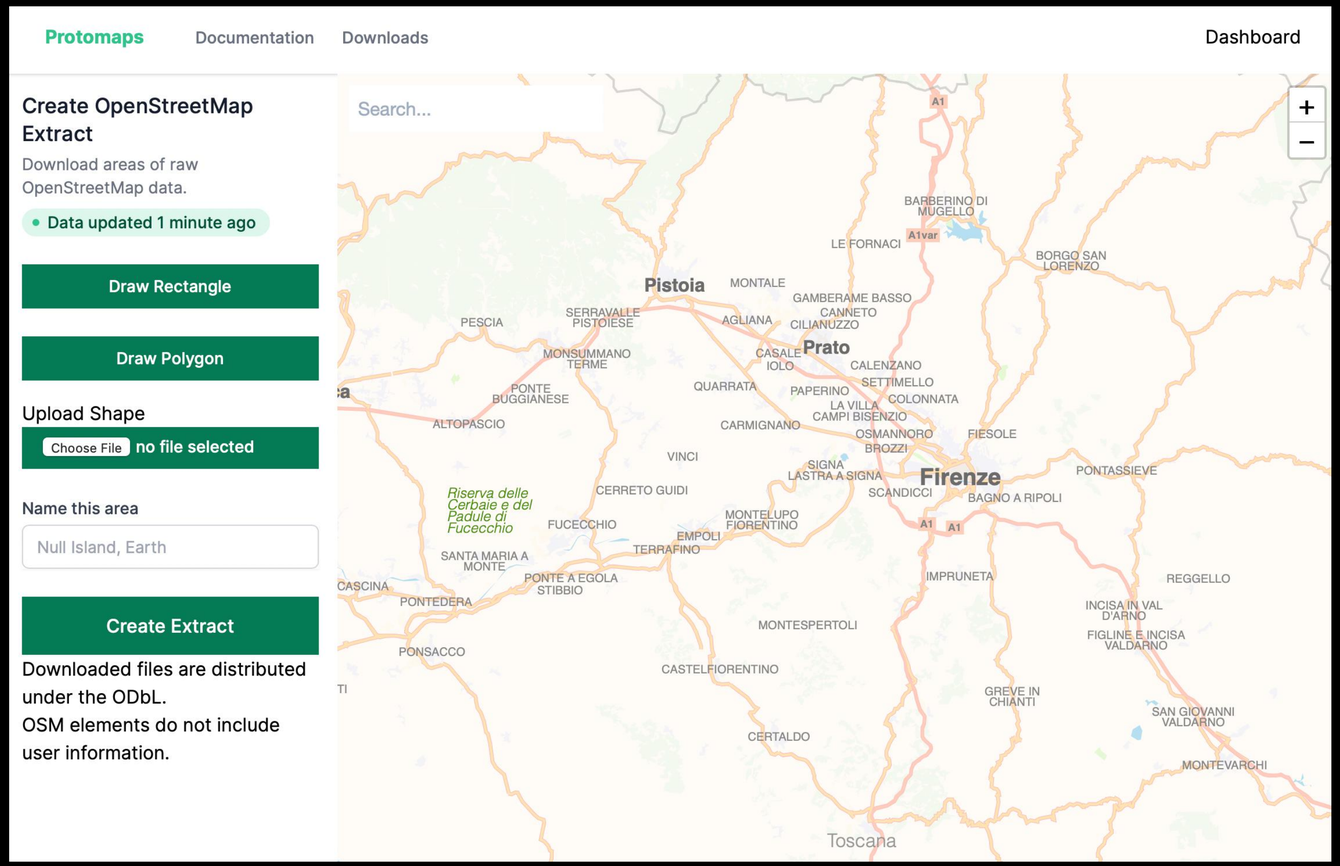
San Francisco

Mapzen

Protomaps Extracts

Free and open source web service at protomaps.com/downloads

- Efficient on-demand metro extracts
- 1-minute update frequency from planet replication feed
- Fast non-rectangular extract areas via S2 Cell indexing
- Current output limited to raw OSM PBF



Create OpenStreetMap Extract

Download areas of raw OpenStreetMap data.

Data updated 1 minute ago

Draw Rectangle

Draw Polygon

Upload Shape

Choose File no file selected

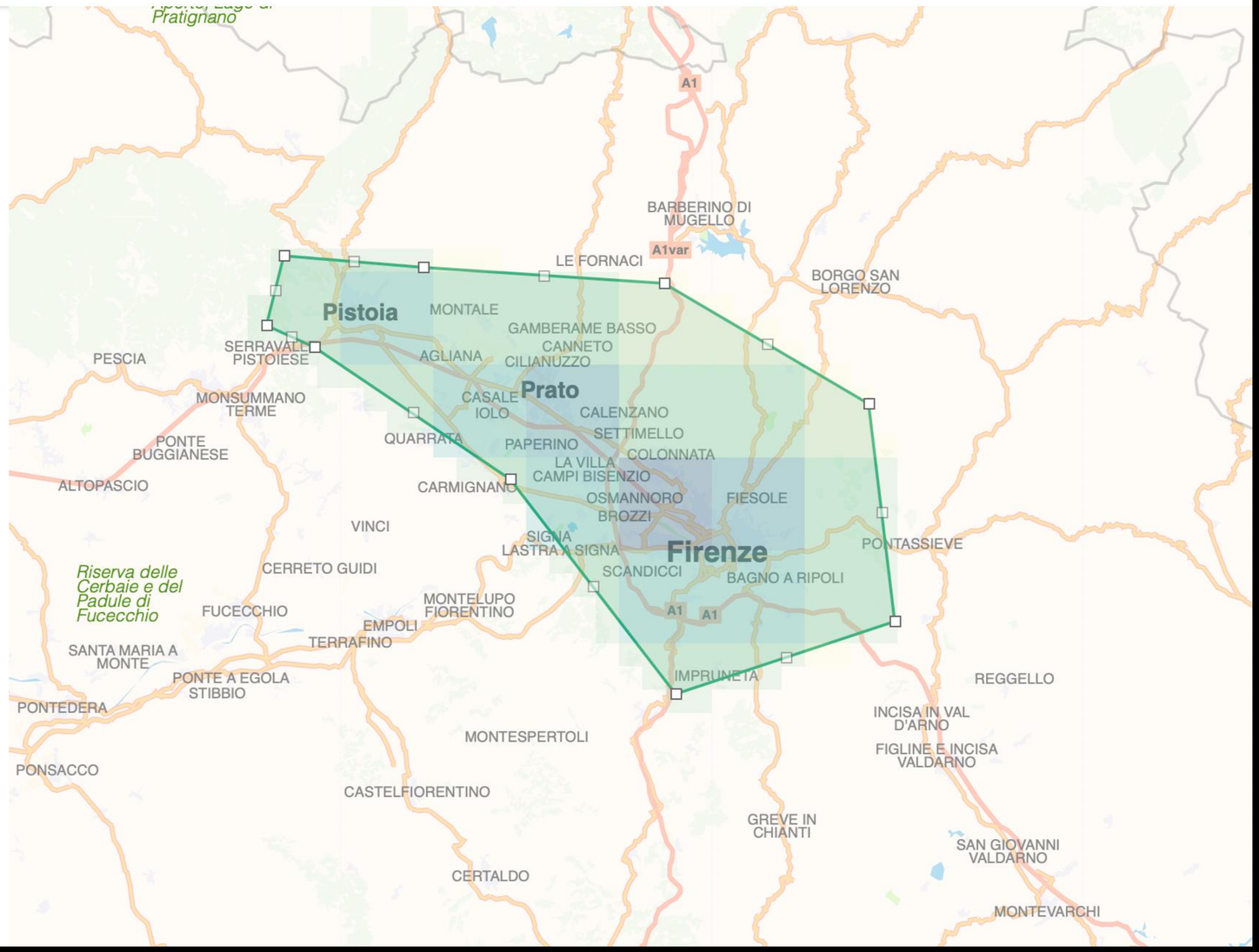
Selected area

This area contains about 1,575,496 nodes, which will take approximately 0.7 minutes.

Name this area

Null Island, Earth

Create Extract



← Back to create new

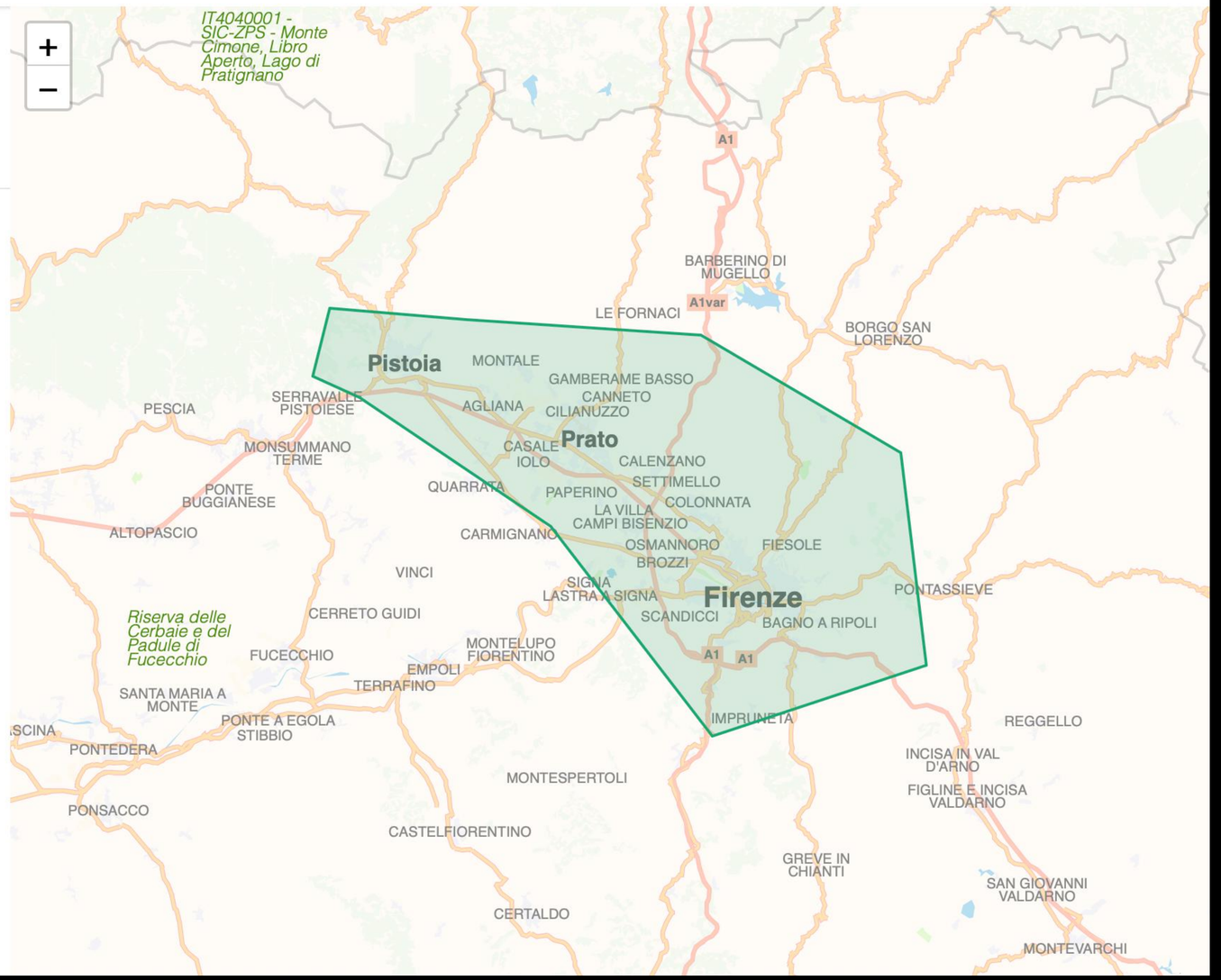
Download OpenStreetMap Extract

706 / 706 cells read.

1515624 / 1722264 nodes read.

0 / 0 elements written.

Boundary GeoJSON



.OSM.PBF

Why not use use these everywhere?

Problem #1: OSM Isn’t Tabular

- OSM is made of nodes, ways, and relations to aid in editing and versioning
- **But,** GIS users expect a tabular schema with points, linestrings and polygons
  - Popular software like QGIS will try very hard to automate this, and not always succeed
  - Can slowly auto-generate thousands of columns where most column values are blank

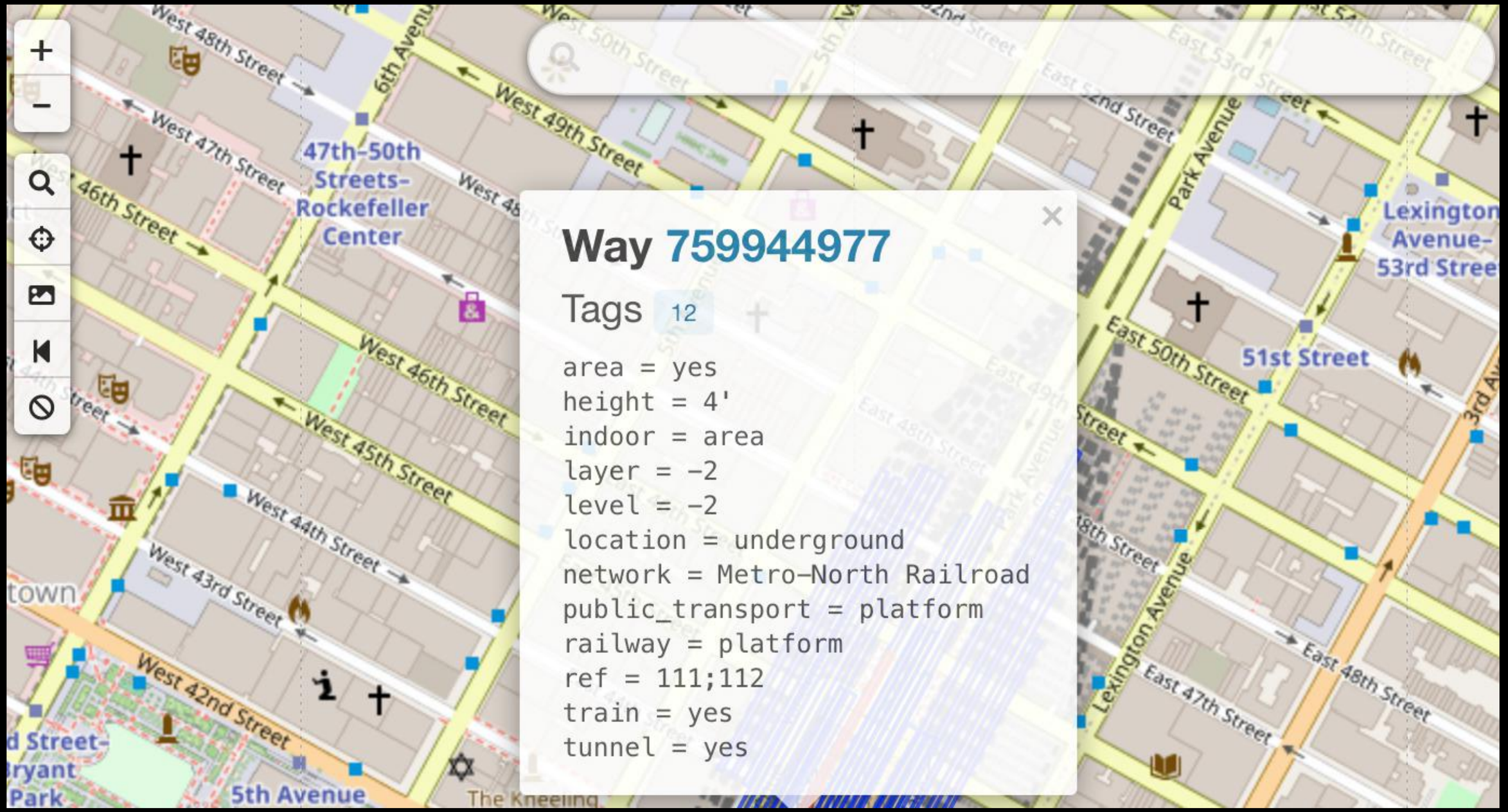
Problem #2: OSM Data can exceed format limits

- The most widely used GIS format is ESRI Shapefile
- Shapefiles have a max size of 4 GB, limit of 255 columns with 10 character column names
- **switchfromshapefile.org**

Problem #3: Making OSM tabular implies lots of cleanup

- `height: 100 m`
- `height: 300’`
- `height: 30.5`
- A strictly typed `number` column needs to handle all of these!

Height	Tagging	Remark
Four metres	`height=4`	Specifying the units is optional for metres, and would look like: `height=4 m`.
1.35 metres	`height=1.35`	Use the decimal / full stop / period character as a separator.
7 feet and 4 inches	`height=7'4"`	Use a typewriter apostrophe character after the feet value and a typewriter double-quote character after the inches value. Spaces are not permitted.


Way 759944977
Tags 12
area = yes
height = 4'
indoor = area
layer = -2
level = -2
location = underground
network = Metro-North Railroad
public_transport = platform
railway = platform
ref = 111;112
train = yes
tunnel = yes
West 48th Street
West 47th Street
West 46th Street
West 45th Street
West 44th Street
West 43rd Street
West 42nd Street
West 49th Street
6th Avenue
47th-50th Streets-Rockefeller Center
5th Avenue
Park Avenue
East 50th Street
51st Street
Lexington Avenue
East 48th Street
East 47th Street
Lexington Avenue-53rd Stree
The Kneeling

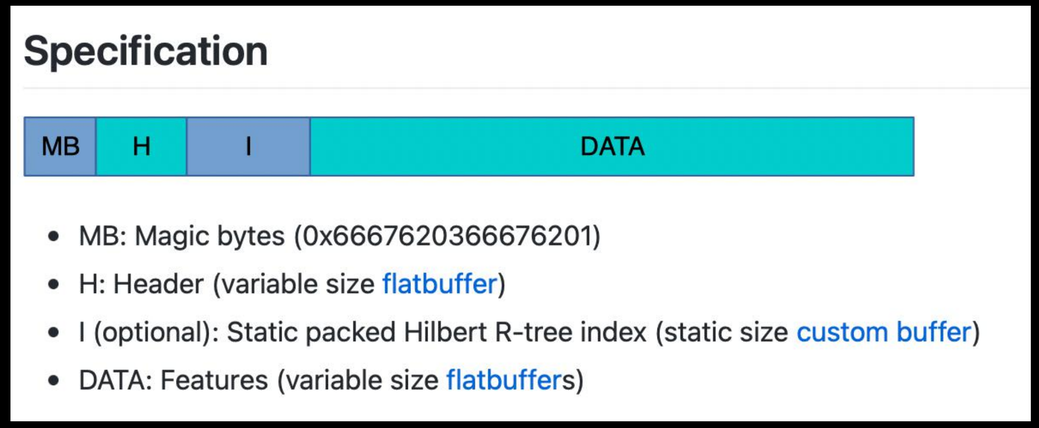
Making OSM tabular

- Status quo approaches:
  - Shapefile, GeoPackage, GDAL OSM Driver, GeoJSON
- Usually, getting *“all OSM data”* into tabular form is not feasible
  - Focus instead on a thematic slice of data

FlatGeobuf

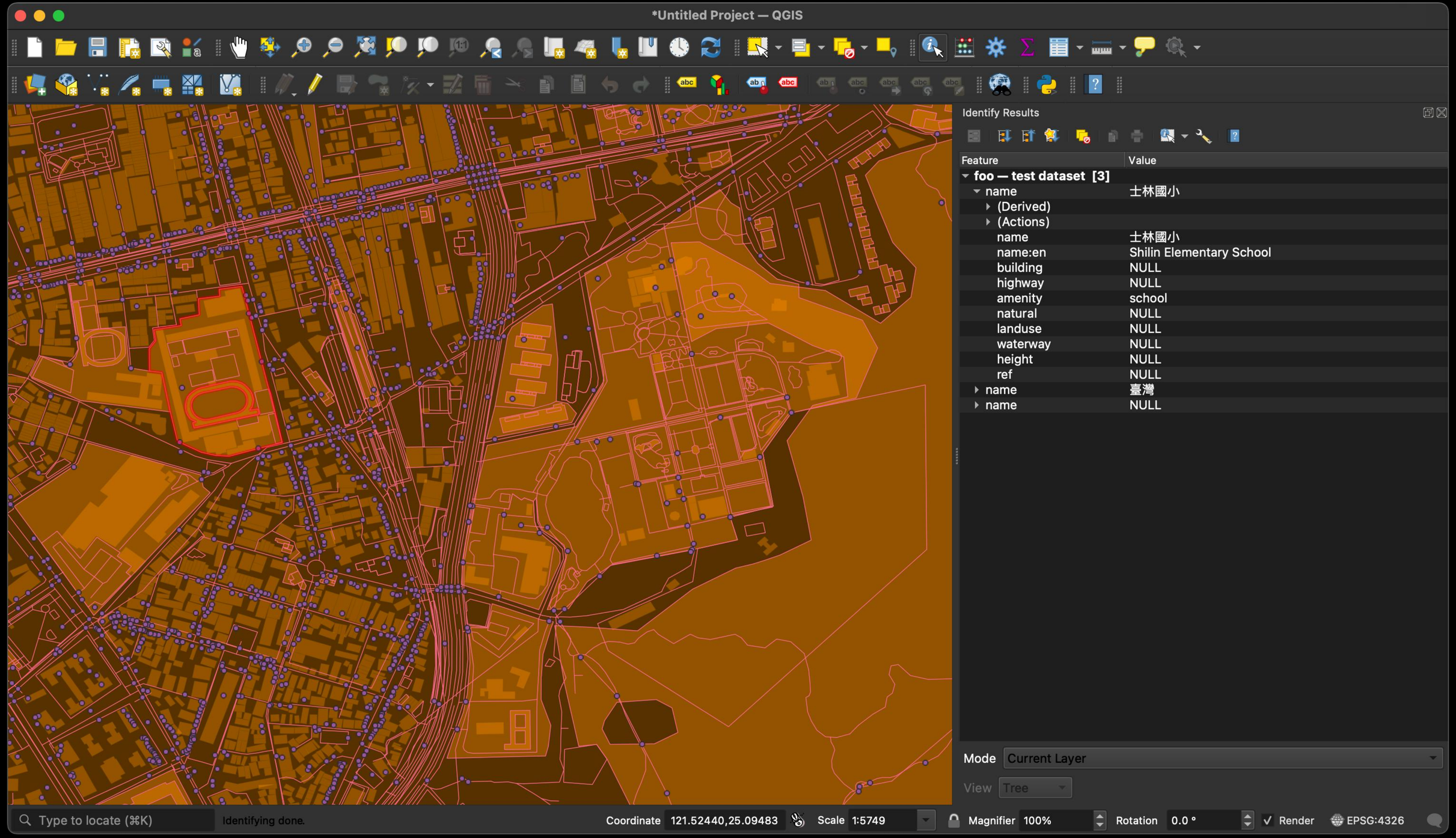
An efficient replacement for Shapefile

- The best current format for general-purpose Simple Features
  - Readable by GDAL 3.1+, QGIS (fixed tabular schema, no schemaless yet)
  - Built support into tippecanoe for reading
    - github.com/protomaps/tippecanoe
- Other alternatives: GeoParquet?



osmx-featuresource

- New OSM Express sub-program to output FlatGeobuf to a stream
  - Uses OSM Express (.osmx) file as a data source
  - Uses libosmium MultiPolygon builder
  - Limited to fixed schema, customization in development
  - github.com/protomaps/OSMExpress/
- `osmx-featuresource SOURCE.osmx MIN_X MIN_Y MAX_X MAX_Y`


*Untitled Project — QGIS
Identify Results
Feature
Value
foo — test dataset [3]
name
士林國小
(Derived)
(Actions)
name
士林國小
name:en
Shilin Elementary School
building
NULL
highway
NULL
amenity
school
natural
NULL
landuse
NULL
waterway
NULL
height
NULL
ref
NULL
name
臺灣
name
NULL
Mode
Current Layer
View
Tree
Type to locate (⌘K)
Identifying done.
Coordinate
121.52440,25.09483
Scale
1:5749
Magnifier
100%
Rotation
0.0 °
Render
EPSG:4326

transformation config approaches

- JSON configuration
  - Mapbox/MapLibre GL styles, osmium-tool config files
- YAML configuration
  - HOT Export Tool, Tangram styles, planetiler custommap (WIP)
- Custom configuration file:
  - osm2pgsql
- Scripting: Tilemaker Lua, Starlark
  - See Cartographic Scripting @ github.com/bdon/protomaps-experiments

Specialized OSM datasets?
e.g. admin boundaries, coastlines

Other OSM-specific datasets might need special techniques

Protoshapes

Planet-scale administrative boundary dataset

- Open source at GitHub.com/protomaps/protoshapes
- A “polygon gazetteer”, or set of polygon shapes with names
- Is not itself a geocoder - but essential input data to enable geocoding and reverse geocoding

OpenStreetMap | Edit | History | Export | GPS Traces | User Diaries | Copyright | Help | About | bdon

Search | Where is this? | Go

Relation: Florence (42602)

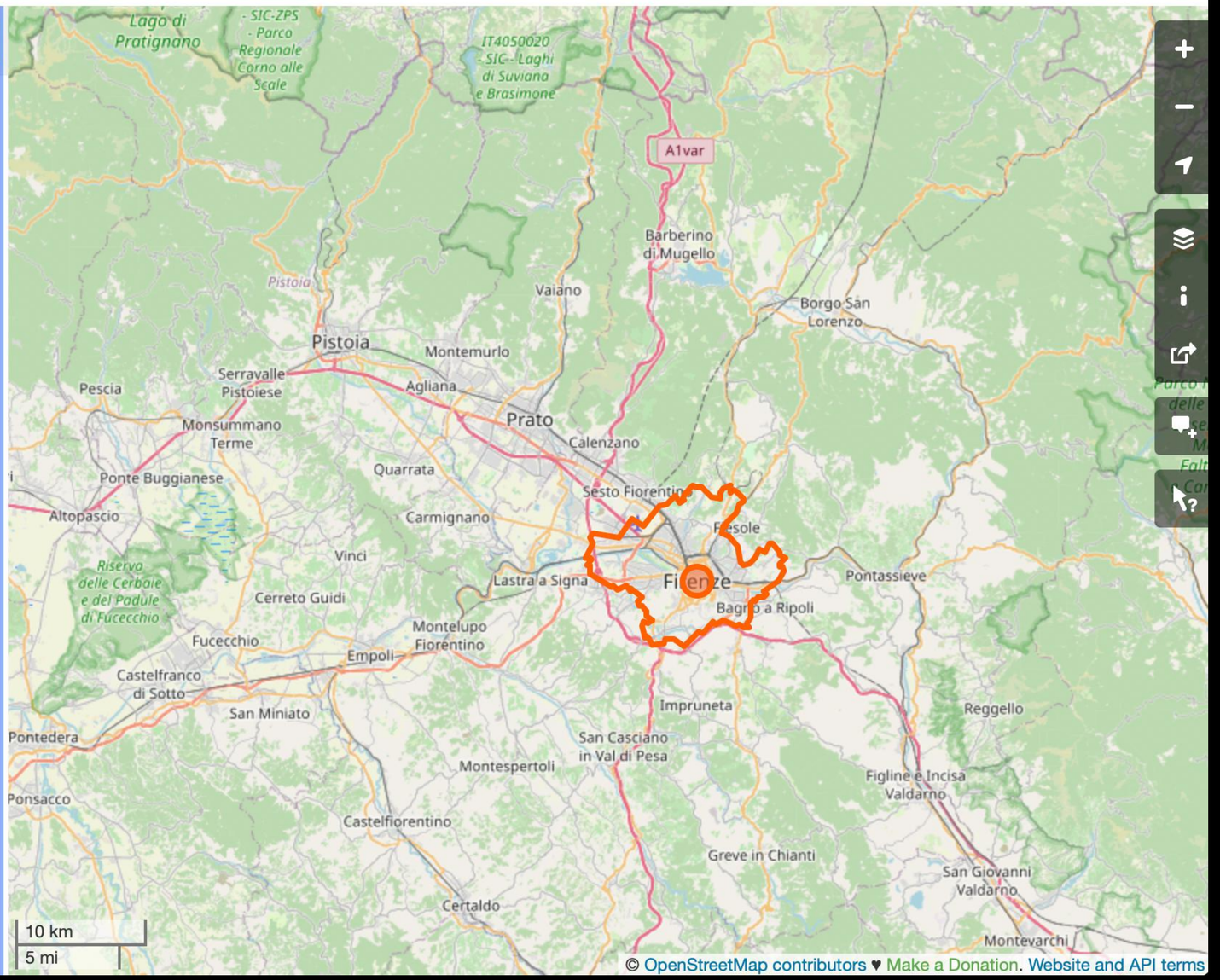
Version #31

Firenze - added slovak name

Edited 9 days ago by *Martin*
Changeset #124732938

Tags

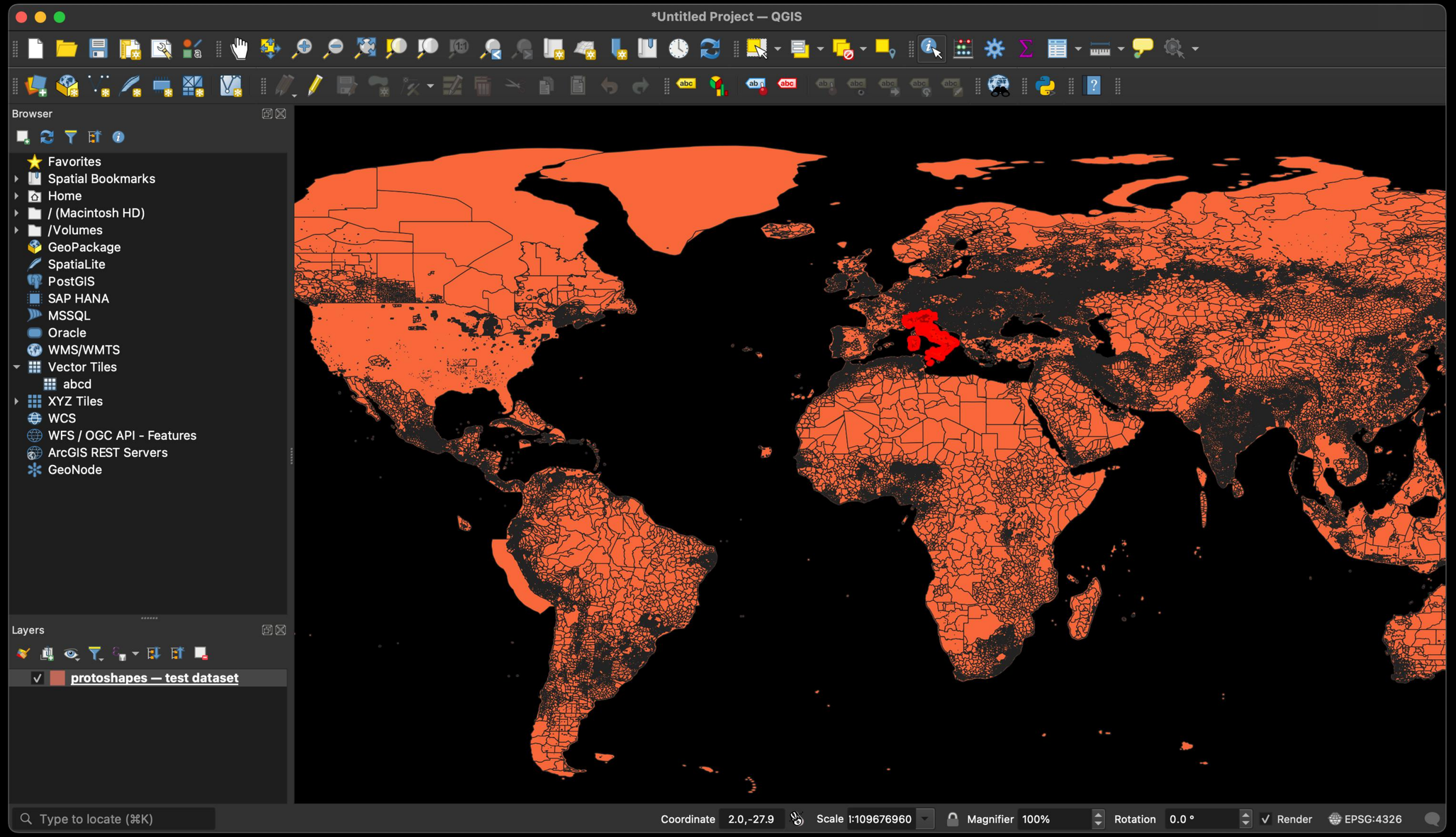
admin_level	8
alt_name:gl	Florenza
boundary	administrative
name	Firenze
name:ca	Florència
name:en	Florence
name:es	Florencia
name:fr	Florence
name:gl	Florencia
name:it	Firenze

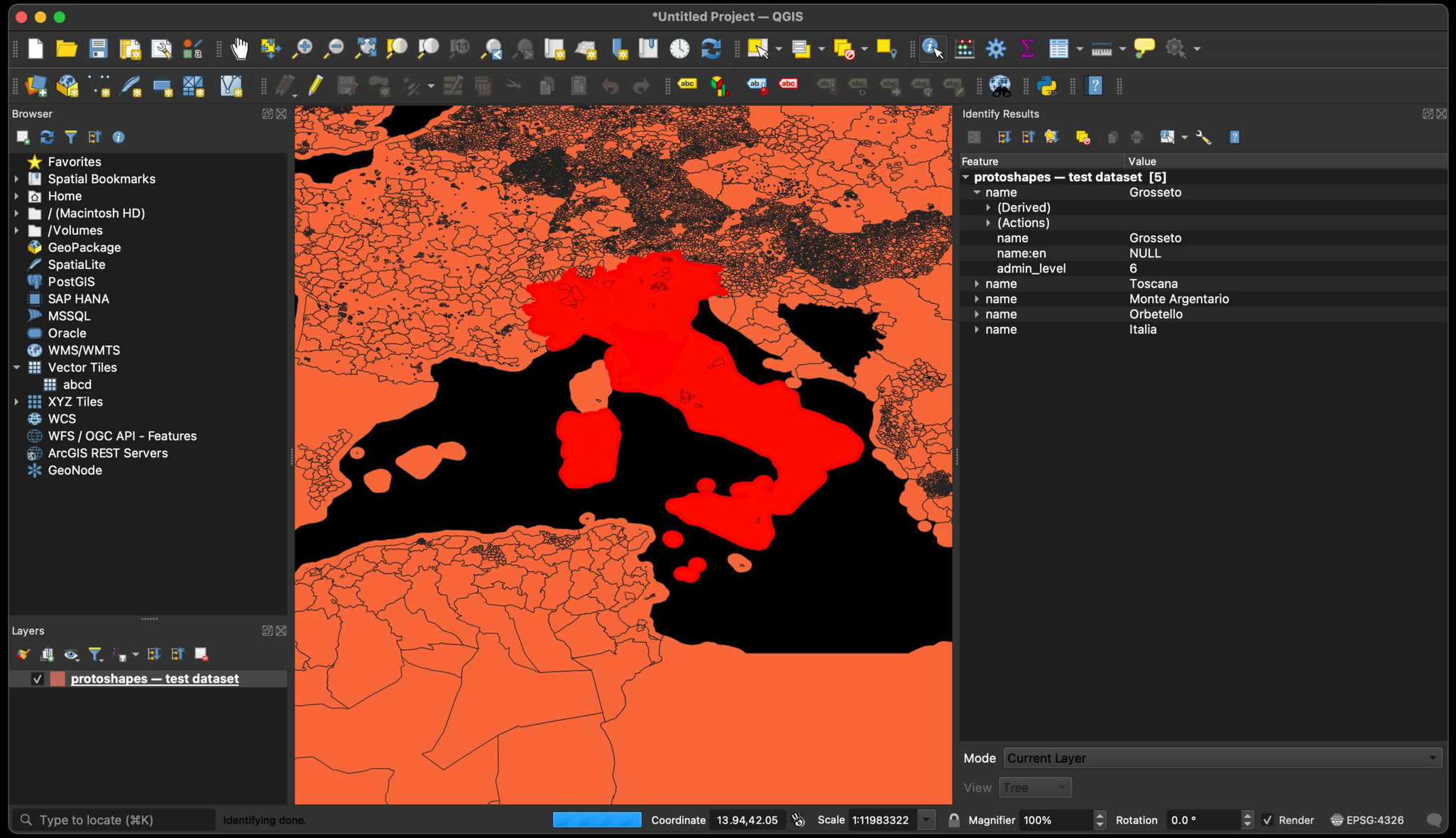


Protoshapes

Another OSM Express-based program

- `protoshapes create planet.osmx protoshapes.fgb`
  - Takes ~30 minutes to output FlatGeobuf on laptop
  - Random access over OSM data useful for other tasks
    - e.g. calculate the degree of road connectivity for a given node
- 4.76 GB, 584,976 polygon geometries
- Does not yet implement spatial hierarchy e.g.
  - `Firenze (6) / Toscana (4) / Italia (2)`

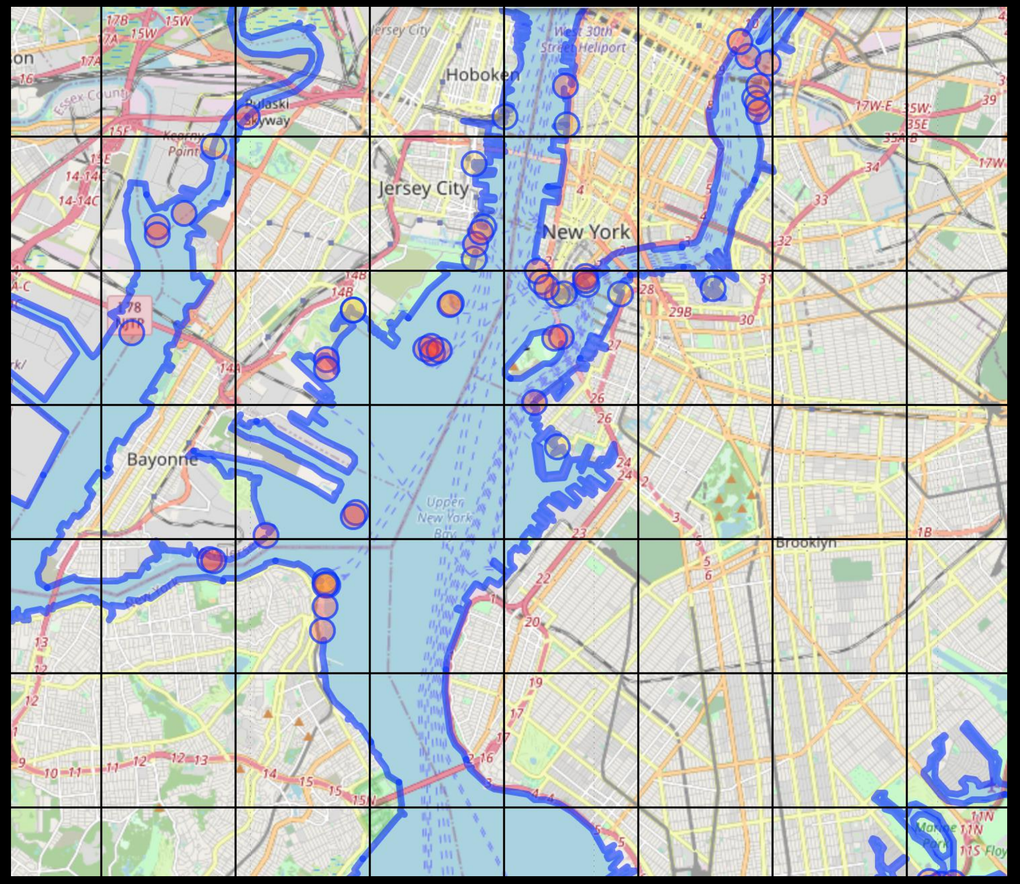

*Untitled Project — QGIS
Browser
Favorites
Spatial Bookmarks
Home
/ (Macintosh HD)
/Volumes
GeoPackage
SpatiaLite
PostGIS
SAP HANA
MSSQL
Oracle
WMS/WMTS
Vector Tiles
abcd
XYZ Tiles
WCS
WFS / OGC API - Features
ArcGIS REST Servers
GeoNode
Layers
protoshapes — test dataset
Type to locate (⌘K)
Coordinate 2.0,-27.9
Scale 1:109676960
Magnifier 100%
Rotation 0.0 °
Render
EPSG:4326


*Untitled Project — QGIS
Browser
Favorites
Spatial Bookmarks
Home
/ (Macintosh HD)
/Volumes
GeoPackage
SpatiaLite
PostGIS
SAP HANA
MSSQL
Oracle
WMS/WMTS
Vector Tiles
abcd
XYZ Tiles
WCS
WFS / OGC API - Features
ArcGIS REST Servers
GeoNode
Layers
protoshapes — test dataset
Identify Results
Feature
Value
protoshapes — test dataset [5]
name
Grosseto
(Derived)
(Actions)
name
Grosseto
name:en
NULL
admin_level
6
name
Toscana
name
Monte Argentario
name
Orbetello
name
Italia
Mode
Current Layer
View
Tree
Type to locate (⌘K)
Identifying done.
Coordinate
13.94,42.05
Scale
1:11983322
Magnifier
100%
Rotation
0.0 °
Render
EPSG:4326

natural=coastline

- It is possible to assemble the ocean and land without dealing with continent-scale polygons!
- Divide the world into a grid and store only the coastline polygons
  - “Web” tiles at a fixed zoom are a convenient starting grid

Hoboken
Jersey City
New York
Bayonne
Brooklyn


Hoboken
Jersey City
New York
Bayonne
Brooklyn

bdon's Diary

Deep Dive: natural=coastline

Posted by bdon on 28 December 2019 in English (English). Last updated on 29 December 2019.

I have made several edits around the world related to the OSM coastline. My goal is to enable small-scale derivation of land and ocean polygons without resorting to global preprocessed continent geometries assembled from programs like OSMCoastline.

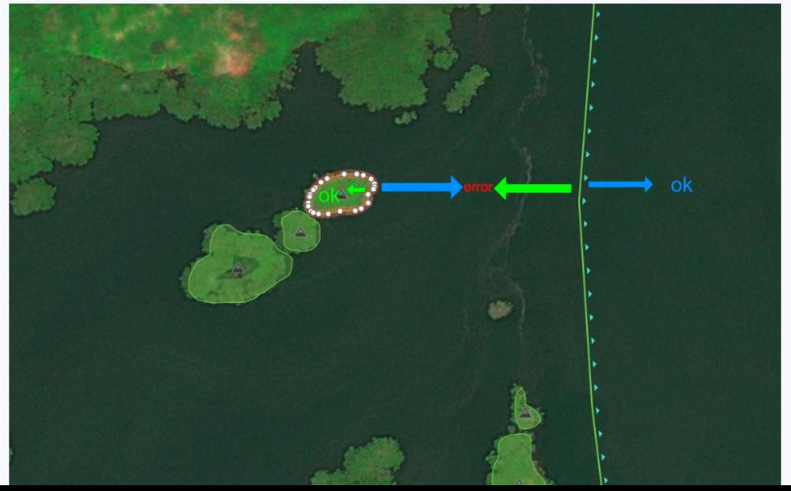
As a primer, the coastline should be mapped as **ways with natural=coastline, with land on the left side and ocean on the right side**. This is specified on the OSM wiki: Tag:natural=coastline. “Land” in this instance is defined as the non-ocean parts of the world, not solid ground in general; for example, the Great Lakes are represented not by coastlines but as water body features inside “land”.

There are a few implications to this design:

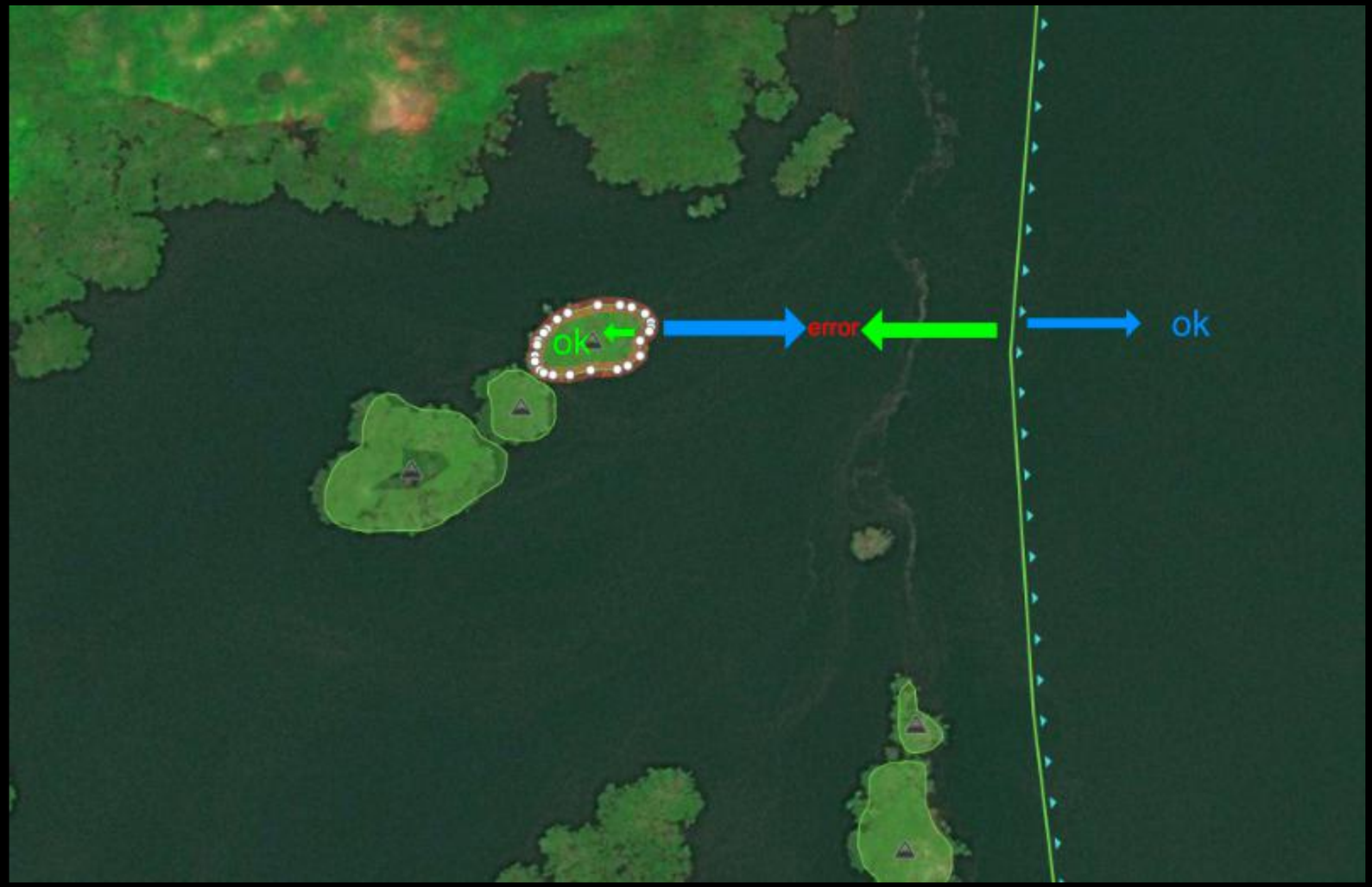
The ocean should be **one polygon** in the OGC Simple Features sense: it has **one outer ring in the clockwise (CW) direction**, and a counter-clockwise (CCW) inner ring for every continent and island. The Caspian Sea is the one exception to this single polygon as stated on the Wiki.

The complement of the global ocean polygon is thousands of land polygons representing continents and islands. Each polygon has a **single CCW outer ring** and zero inner rings. Again, the one exception is that the Eurasian continent polygon has a CW inner ring representing the Caspian Sea.

In theory, this specification should be enough to infer land and ocean polygons from orientation even within a small geographic extract. I discovered and corrected several dozen places where a violation of this specification arises. The image below is an example of data that needs to be corrected:



natural=coastline

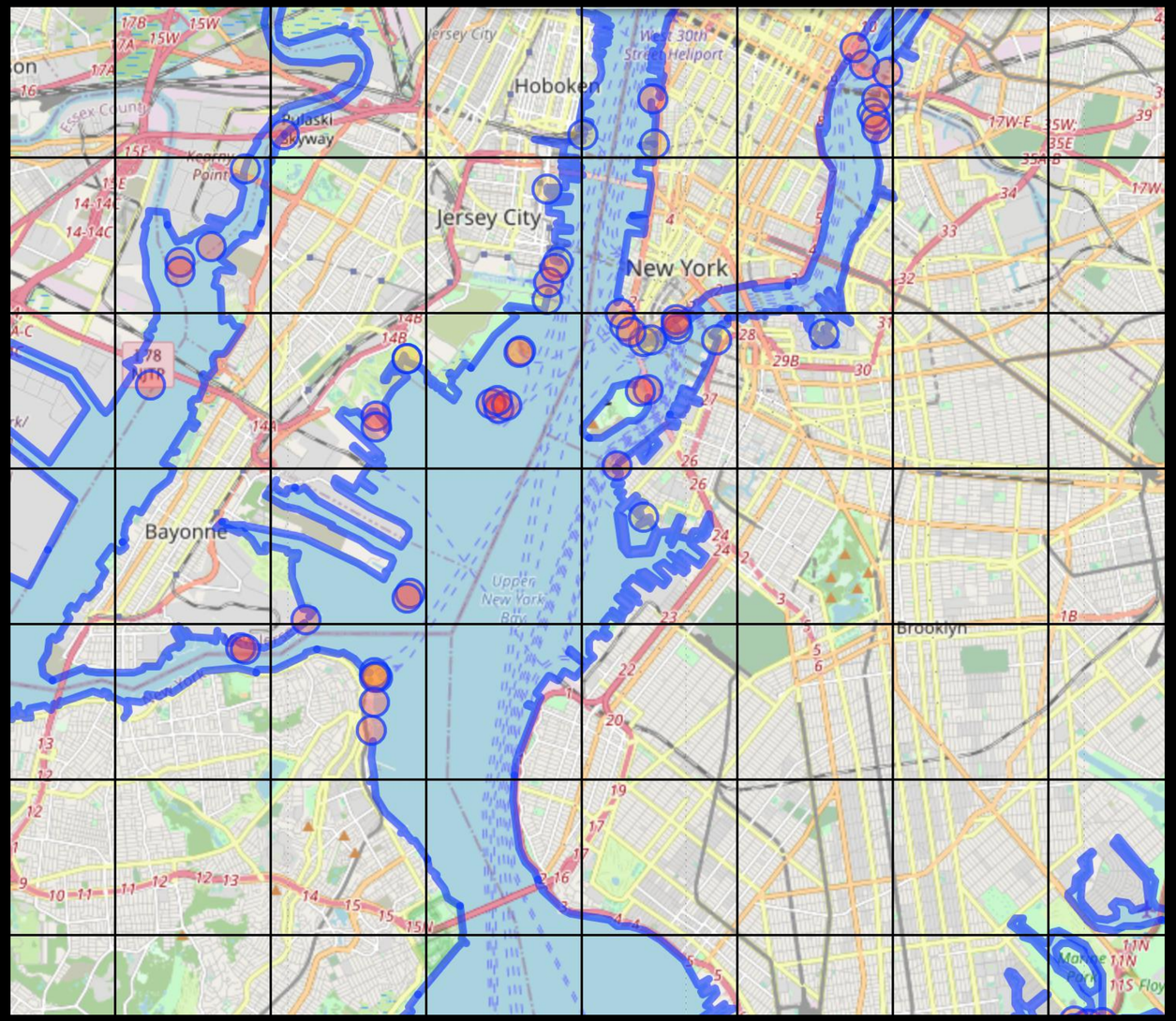


Carcassonne

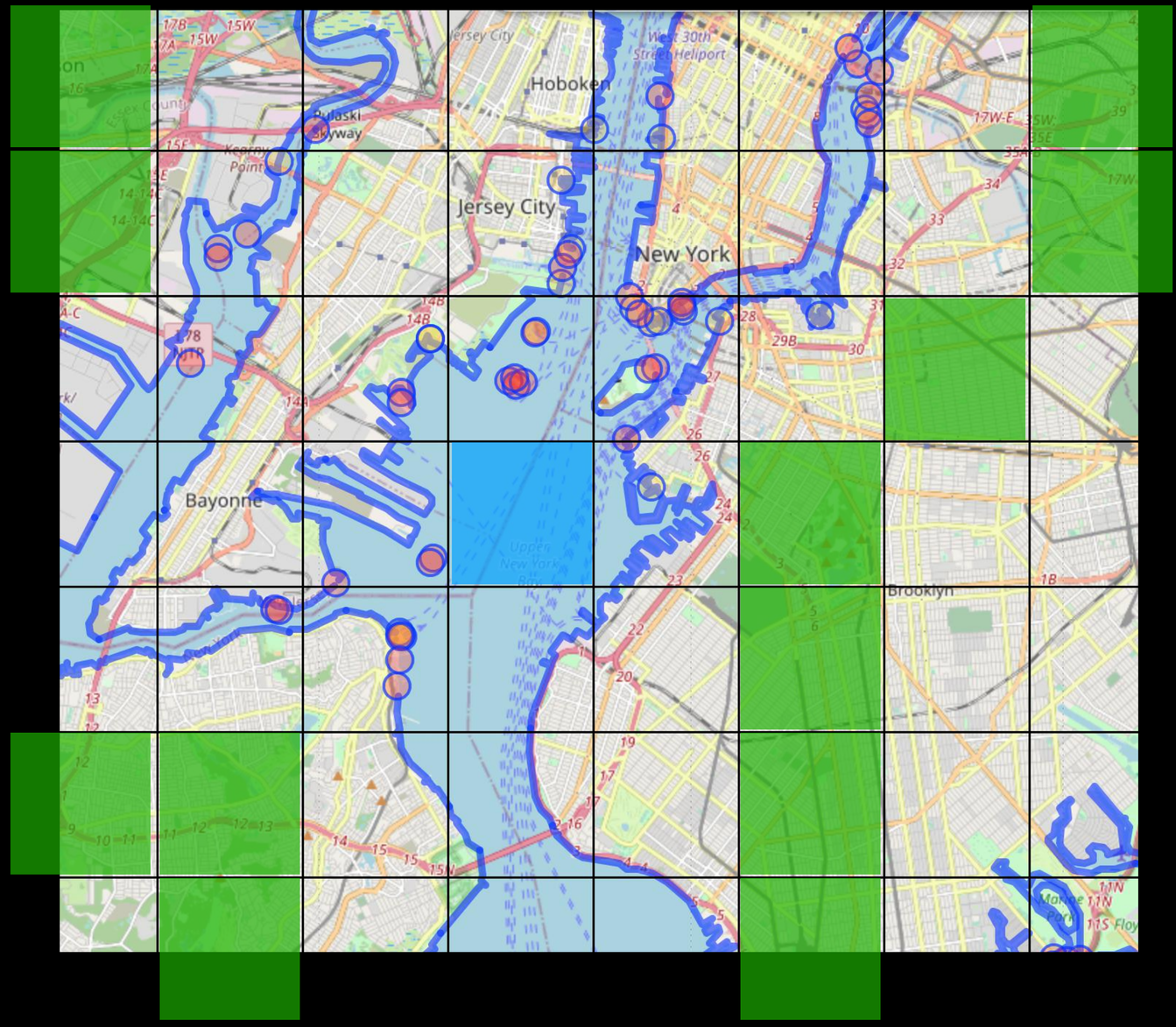
Source © Z-Man Games

Koefbac, CC BY-SA 4.0 <https://creativecommons.org/licenses/by-sa/4.0>, via Wikimedia Commons

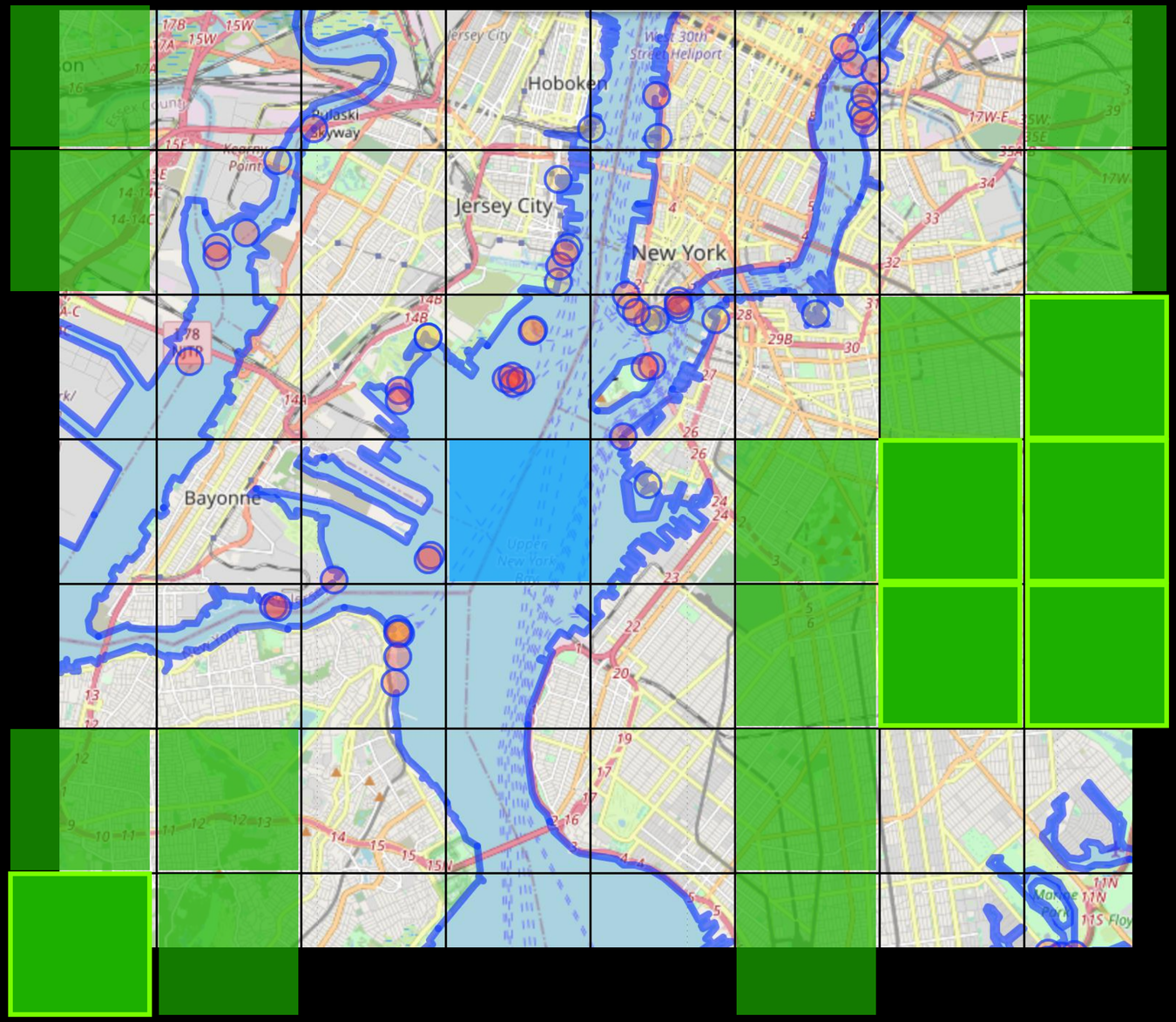
Apply flood-fill to derive land and water



Apply flood-fill to derive land and water



Apply flood-fill to derive land and water

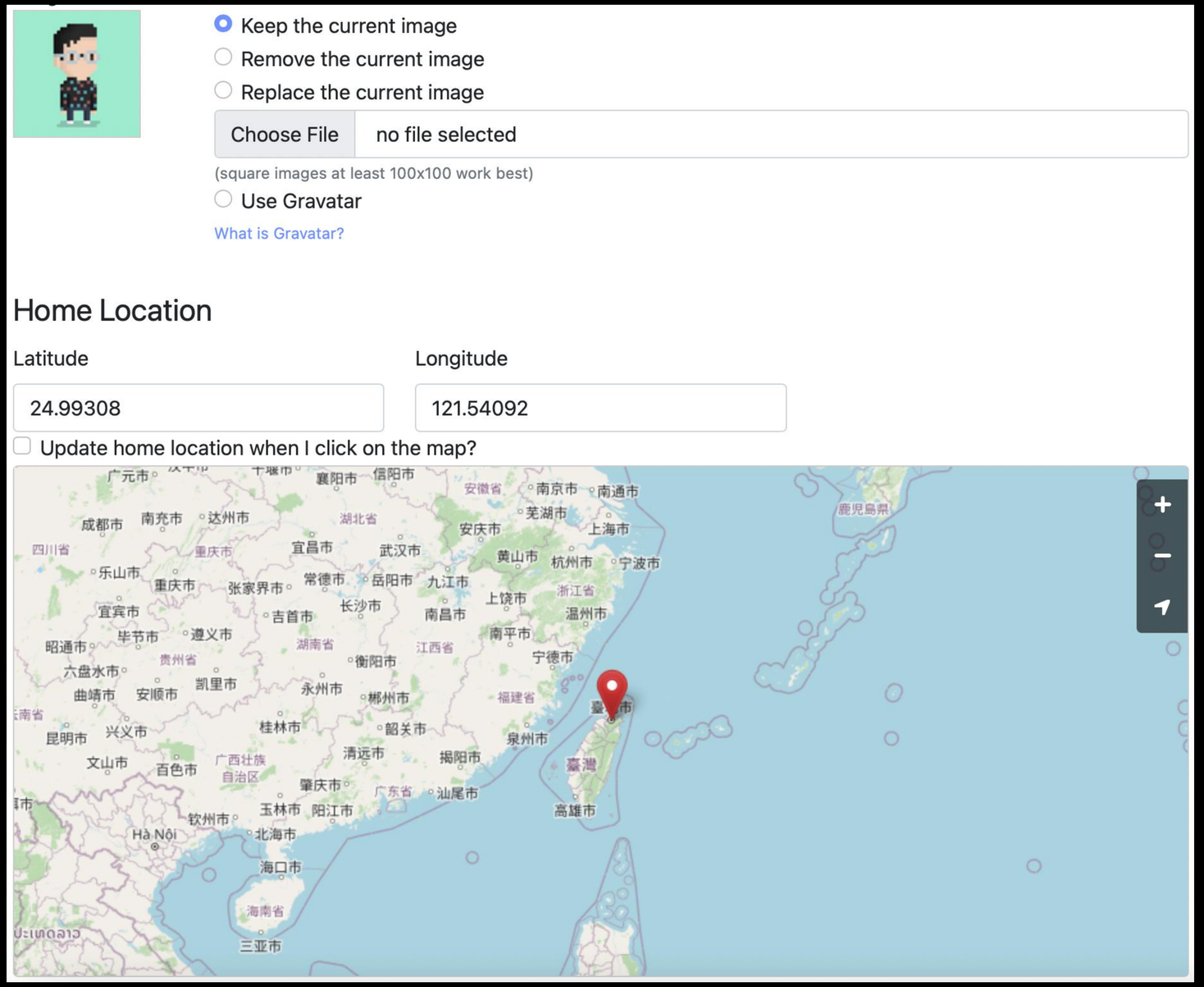


Store ocean=true boolean as 1-bit image

Conclusions

Or why any of this matters

- **“Raw” OpenStreetMap** might always be exotic
  - To bring OSM to widest audience, a bridge to GIS is essential
- Users focus on a specific area of interest + theme, enable **data slicing**
- Make the specialized tools & formats **open source, simple and fast**

Thank you!

- OSM Username: @bdon
- GitHub.com/protomaps
- brandon@protomaps.com