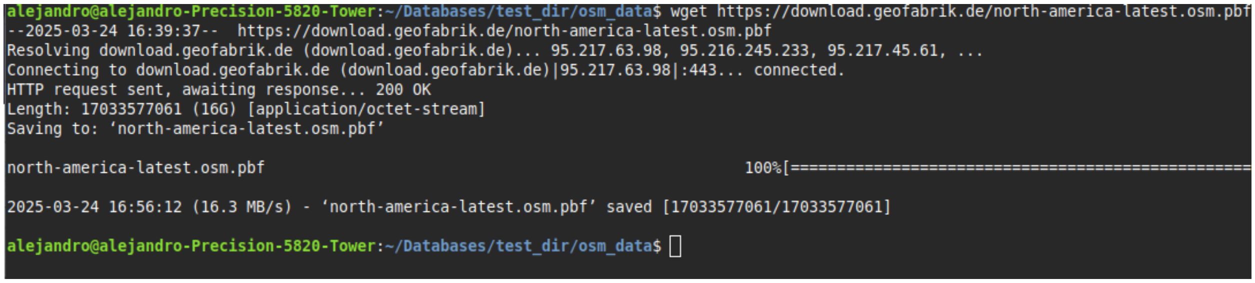
Guide on Downloading OSM Data for Wang et al. Implementation

Wang et al. used up to 151 GB of uncompressed OSM data for the non-larger dataset. According to the OSM, the current planet consists of more than 2 TB of uncompressed data. Our PCs can not handle such a high quantity of data, so we aimed to download a region instead. Geofabrik allows users to download up to date regional OSM data, and to acquire (about) the same amount of data as Wang et al. used , we downloaded the latest north america dataset as a .osm.pbf file from Geofabrik using wget:

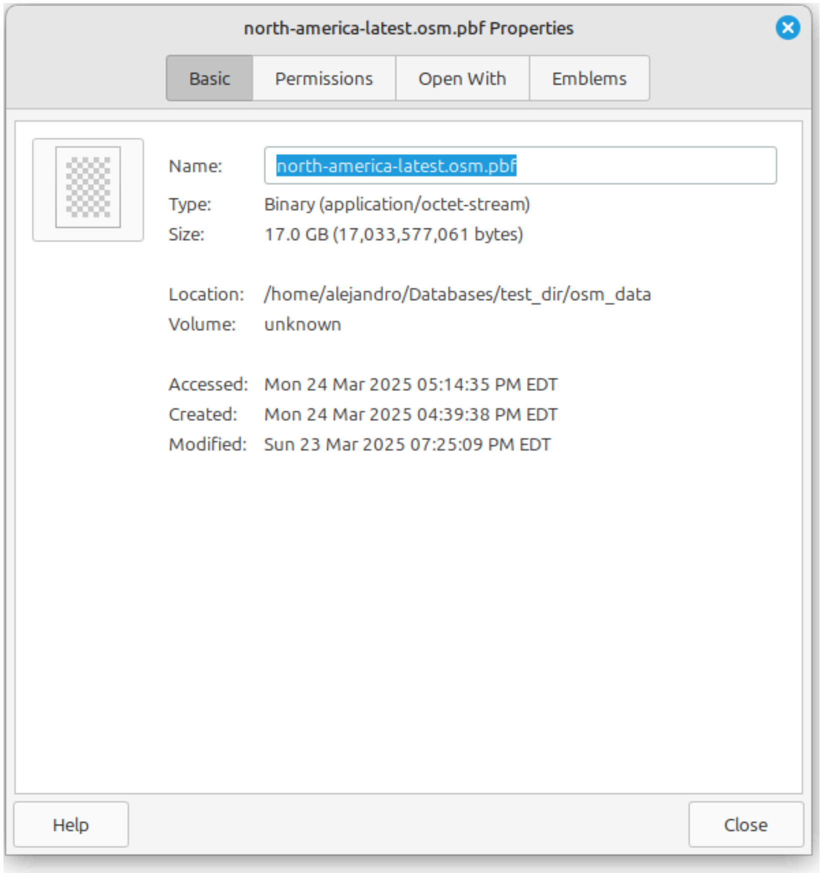


```
alejandro@alejandro-Precision-5820-Tower:~/Databases/test_dir/osm_data$ wget https://download.geofabrik.de/north-america-latest.osm.pbf
--2025-03-24 16:39:37--  https://download.geofabrik.de/north-america-latest.osm.pbf
Resolving download.geofabrik.de (download.geofabrik.de)... 95.217.63.98, 95.216.245.233, 95.217.45.61, ...
Connecting to download.geofabrik.de (download.geofabrik.de)|95.217.63.98|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 17033577061 (16G) [application/octet-stream]
Saving to: ‘north-america-latest.osm.pbf’

north-america-latest.osm.pbf                                        100%[==================================================

2025-03-24 16:56:12 (16.3 MB/s) - ‘north-america-latest.osm.pbf’ saved [17033577061/17033577061]

alejandro@alejandro-Precision-5820-Tower:~/Databases/test_dir/osm_data$ 
```



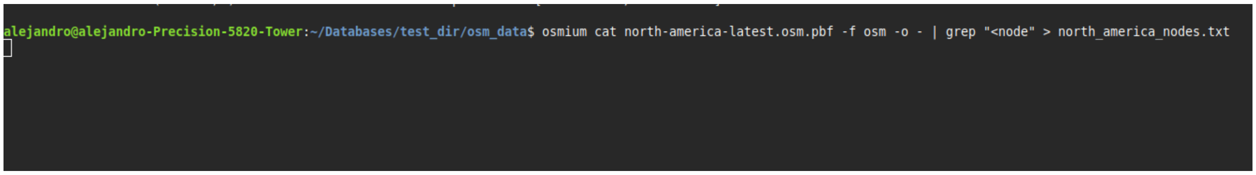
north-america-latest.osm.pbf Properties

Name: north-america-latest.osm.pbf
Type: Binary (application/octet-stream)
Size: 17.0 GB (17,033,577,061 bytes)

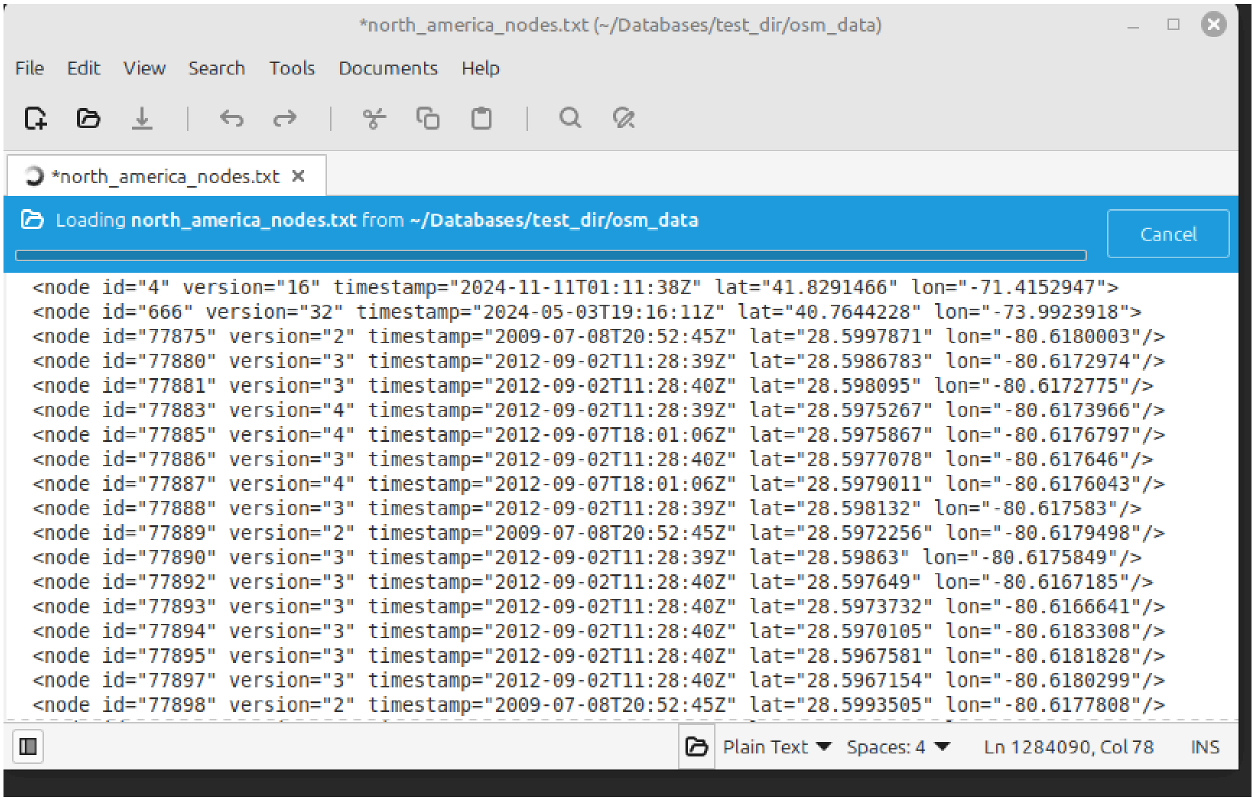
Location: /home/alejandro/Databases/test_dir/osm_data
Volume: unknown

Accessed: Mon 24 Mar 2025 05:14:35 PM EDT
Created: Mon 24 Mar 2025 04:39:38 PM EDT
Modified: Sun 23 Mar 2025 07:25:09 PM EDT

Once downloaded, the node data from the .osm.pbf file had to be converted to a .txt file, which is done with the below command. Also below is what the .txt file looks like, and the node id, version, timestamp, latitude, and longitude fields can be seen in the records with their associated values:

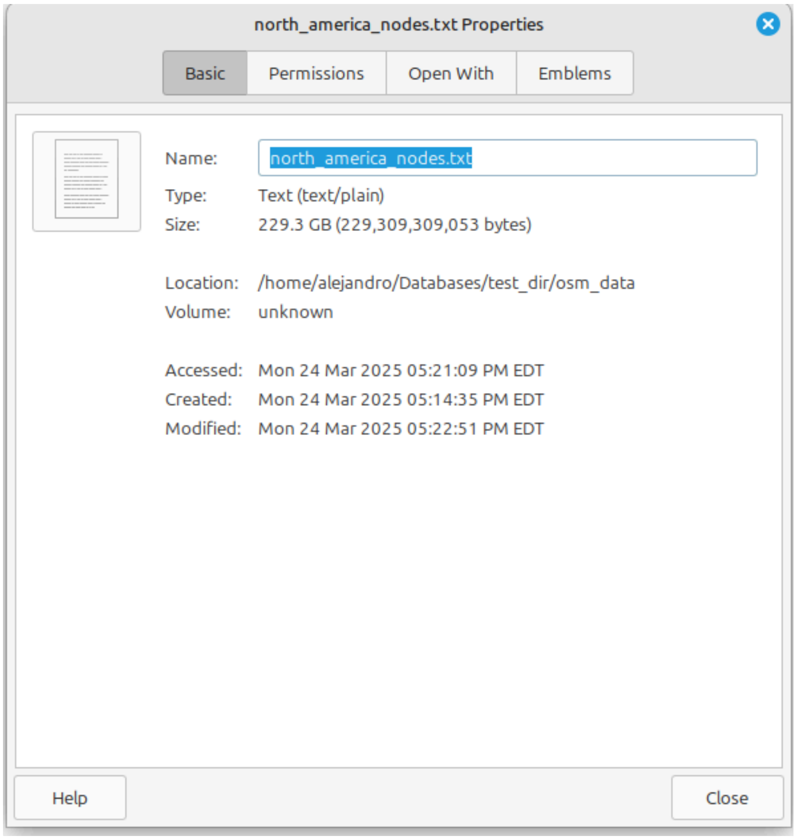


```
alejandro@alejandro-Precision-5820-Tower:~/Databases/test_dir/osm_data$ osmium cat north-america-latest.osm.pbf -f osm -o - | grep "<node" > north_america_nodes.txt
```



Loading **north_america_nodes.txt** from **~/Databases/test_dir/osm_data**

```
  <node id="4" version="16" timestamp="2024-11-11T01:11:38Z" lat="41.8291466" lon="-71.4152947">
  <node id="666" version="32" timestamp="2024-05-03T19:16:11Z" lat="40.7644228" lon="-73.9923918">
  <node id="77875" version="2" timestamp="2009-07-08T20:52:45Z" lat="28.5997871" lon="-80.6180003"/>
  <node id="77880" version="3" timestamp="2012-09-02T11:28:39Z" lat="28.5986783" lon="-80.6172974"/>
  <node id="77881" version="3" timestamp="2012-09-02T11:28:40Z" lat="28.598095" lon="-80.6172775"/>
  <node id="77883" version="4" timestamp="2012-09-02T11:28:39Z" lat="28.5975267" lon="-80.6173966"/>
  <node id="77885" version="4" timestamp="2012-09-07T18:01:06Z" lat="28.5975867" lon="-80.6176797"/>
  <node id="77886" version="3" timestamp="2012-09-02T11:28:40Z" lat="28.5977078" lon="-80.617646"/>
  <node id="77887" version="4" timestamp="2012-09-07T18:01:06Z" lat="28.5979011" lon="-80.6176043"/>
  <node id="77888" version="3" timestamp="2012-09-02T11:28:39Z" lat="28.598132" lon="-80.617583"/>
  <node id="77889" version="2" timestamp="2009-07-08T20:52:45Z" lat="28.5972256" lon="-80.6179498"/>
  <node id="77890" version="3" timestamp="2012-09-02T11:28:39Z" lat="28.59863" lon="-80.6175849"/>
  <node id="77892" version="3" timestamp="2012-09-02T11:28:40Z" lat="28.597649" lon="-80.6167185"/>
  <node id="77893" version="3" timestamp="2012-09-02T11:28:40Z" lat="28.5973732" lon="-80.6166641"/>
  <node id="77894" version="3" timestamp="2012-09-02T11:28:40Z" lat="28.5970105" lon="-80.6183308"/>
  <node id="77895" version="3" timestamp="2012-09-02T11:28:40Z" lat="28.5967581" lon="-80.6181828"/>
  <node id="77897" version="3" timestamp="2012-09-02T11:28:40Z" lat="28.5967154" lon="-80.6180299"/>
  <node id="77898" version="2" timestamp="2009-07-08T20:52:45Z" lat="28.5993505" lon="-80.6177808"/>
```

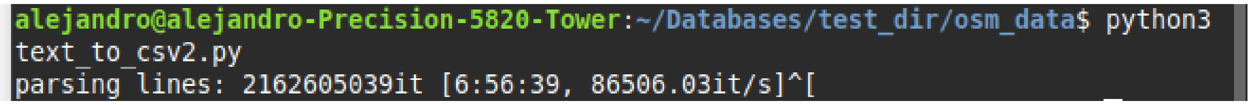


Name: north_america_nodes.txt
Type: Text (text/plain)
Size: 229.3 GB (229,309,309,053 bytes)

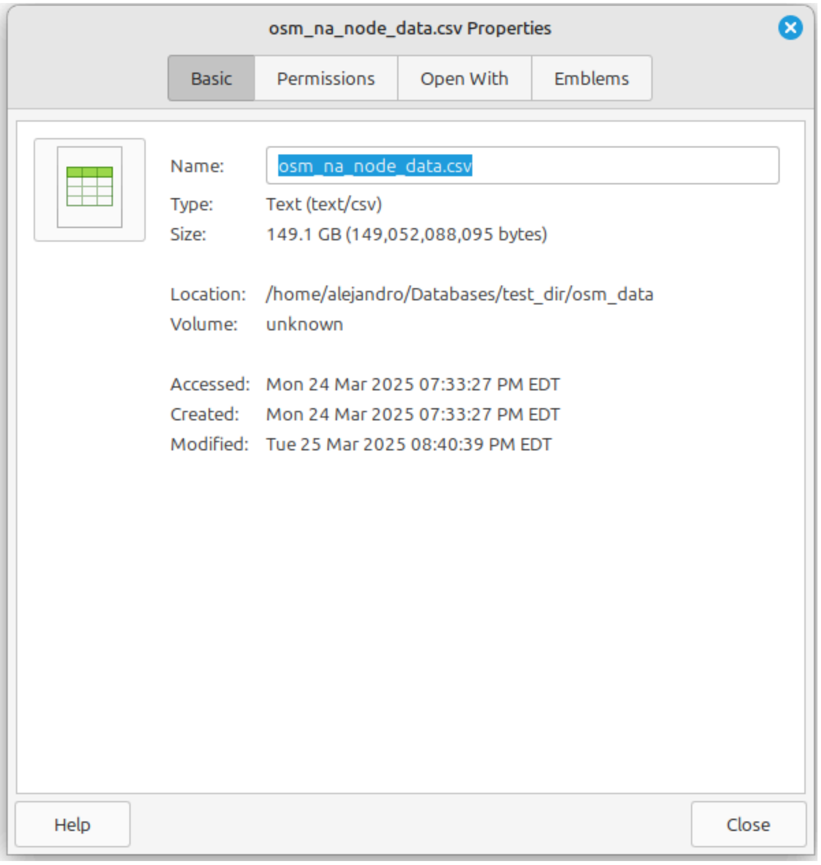
Location: /home/alejandro/Databases/test_dir/osm_data
Volume: unknown

Accessed: Mon 24 Mar 2025 05:21:09 PM EDT
Created: Mon 24 Mar 2025 05:14:35 PM EDT
Modified: Mon 24 Mar 2025 05:22:51 PM EDT

After this, we ran our text_to_cvs2.py converter script to convert the text file into a csv file - removing unnecessary fields, standardizing the timestamp format, and adding the unique 12 byte ID, latitude, and longitude in the process. Total run time was almost 7 hours, and resulted in a .csv file of size 149.1 GB, which is close to the 151 GB used in Wang et al. The resulting .csv file contains all of the data nodes and associated metadata, including the 12 byte ID, longitude, latitude, and timestamp - which are the most relevant for our experimentation.



```
alejandro@alejandro-Precision-5820-Tower:~/Databases/test_dir/osm_data$ python3
text_to_csv2.py
parsing lines: 2162605039it [6:56:39, 86506.03it/s]^[
```



Name: osm_na_node_data.csv

Type: Text (text/csv)

Size: 149.1 GB (149,052,088,095 bytes)

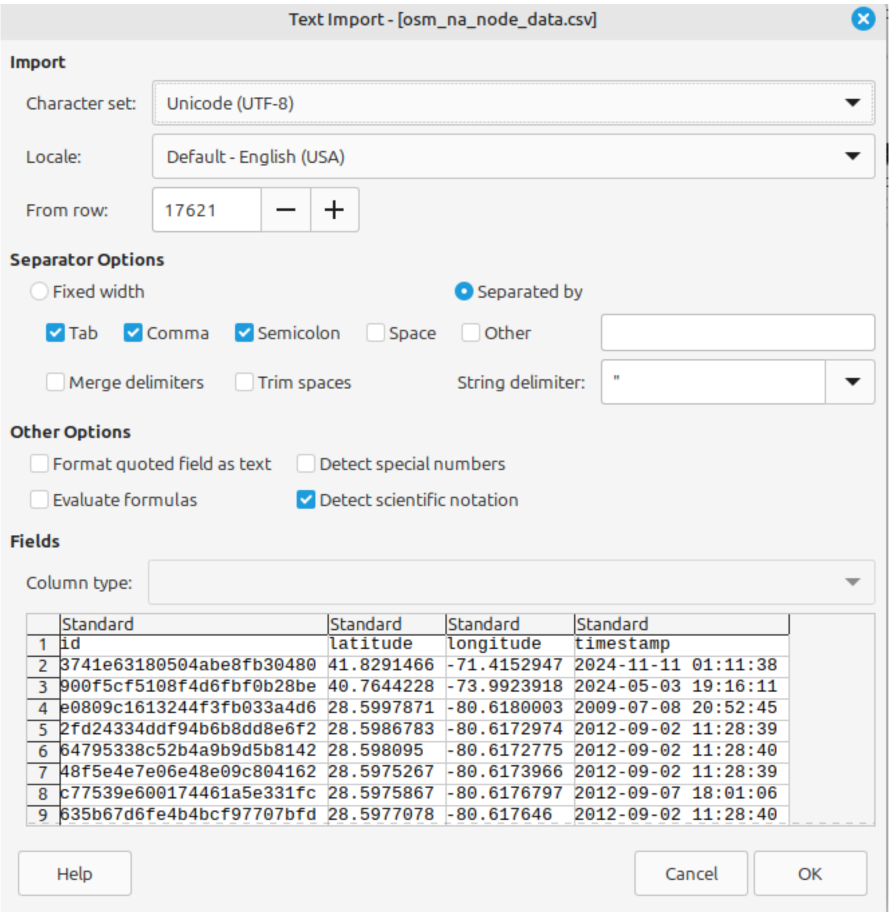
Location: /home/alejandro/Databases/test_dir/osm_data

Volume: unknown

Accessed: Mon 24 Mar 2025 07:33:27 PM EDT

Created: Mon 24 Mar 2025 07:33:27 PM EDT

Modified: Tue 25 Mar 2025 08:40:39 PM EDT



Text Import - [osm_na_node_data.csv]

Import

Character set: Unicode (UTF-8)

Locale: Default - English (USA)

From row: 17621

Separator Options

○ Fixed width ● Separated by

☑ Tab ☑ Comma ☑ Semicolon □ Space □ Other

□ Merge delimiters □ Trim spaces String delimiter: "

Other Options

□ Format quoted field as text □ Detect special numbers

□ Evaluate formulas ☑ Detect scientific notation

Fields

Column type:

	Standard	Standard	Standard	Standard
1	id	latitude	longitude	timestamp
2	3741e63180504abe8fb30480	41.8291466	-71.4152947	2024-11-11 01:11:38
3	900f5cf5108f4d6fbf0b28be	40.7644228	-73.9923918	2024-05-03 19:16:11
4	e0809c1613244f3fb033a4d6	28.5997871	-80.6180003	2009-07-08 20:52:45
5	2fd24334ddf94b6b8dd8e6f2	28.5986783	-80.6172974	2012-09-02 11:28:39
6	64795338c52b4a9b9d5b8142	28.598095	-80.6172775	2012-09-02 11:28:40
7	48f5e4e7e06e48e09c804162	28.5975267	-80.6173966	2012-09-02 11:28:39
8	c77539e600174461a5e331fc	28.5975867	-80.6176797	2012-09-07 18:01:06
9	635b67d6fe4b4bcf97707bfd	28.5977078	-80.617646	2012-09-02 11:28:40

It is important to note that the Geofabrik OSM datasets are constantly exchanging and expanding, as users upload new and updated data. At the time of extraction, the 17 GB osm.pbf North America dataset came out to 149.1 GB in csv format.