**Tag ratios and NER**

For each file, we first ran the Tag ratios algorithm and found out the lines which could contain any measurement data. Then using those results, we read those lines and using the OpenNLP NER and extracted the relevant measurement data that was present in those files. We were able to run this for approximately 500,000 files.

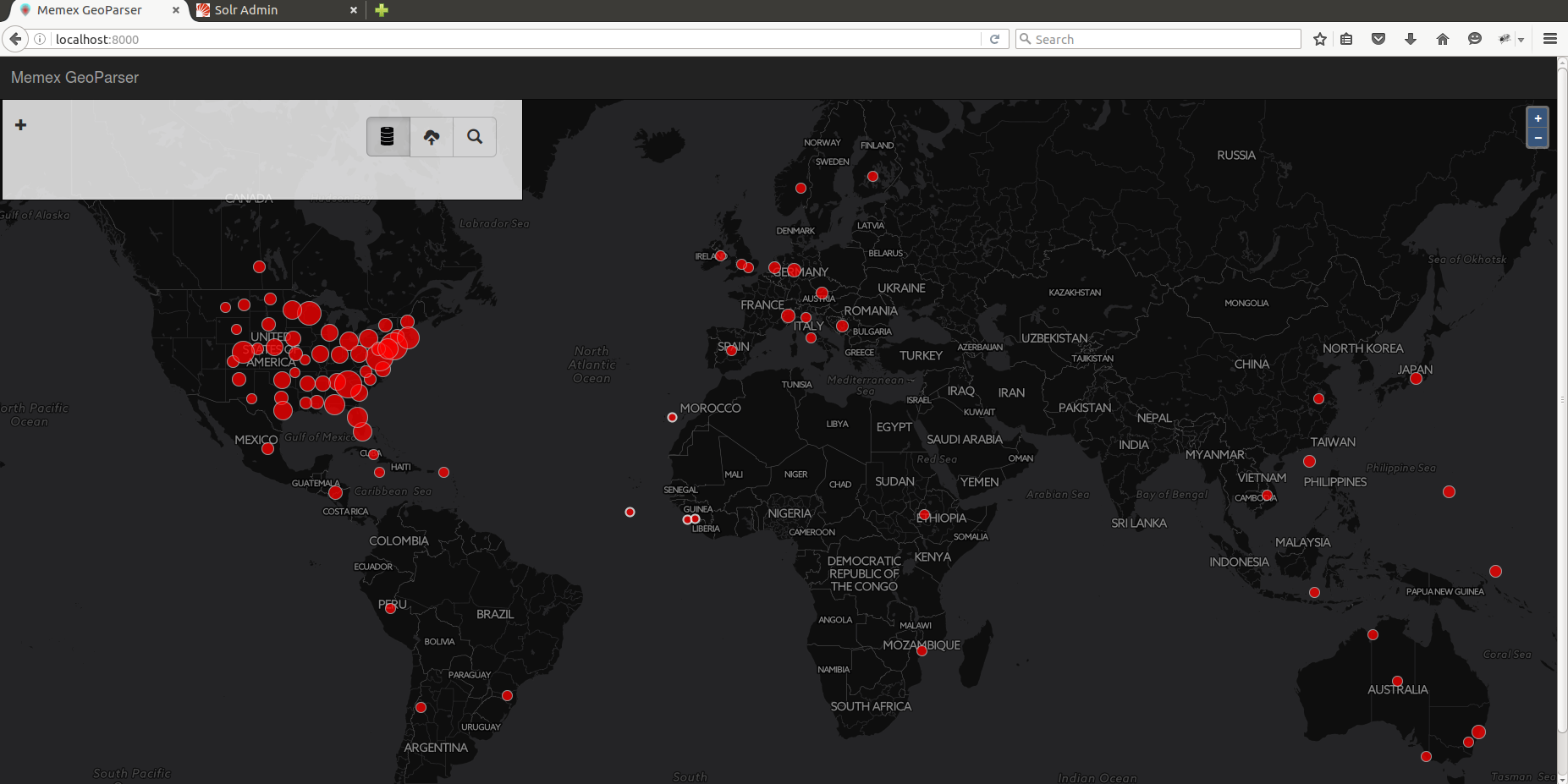
To run the program, we need to point to the correct path where the files are stored. Also, the proper path to the ner.bin files need to be given. The output of this program is a CSV file with the format *(filename, measurement\_data)*.

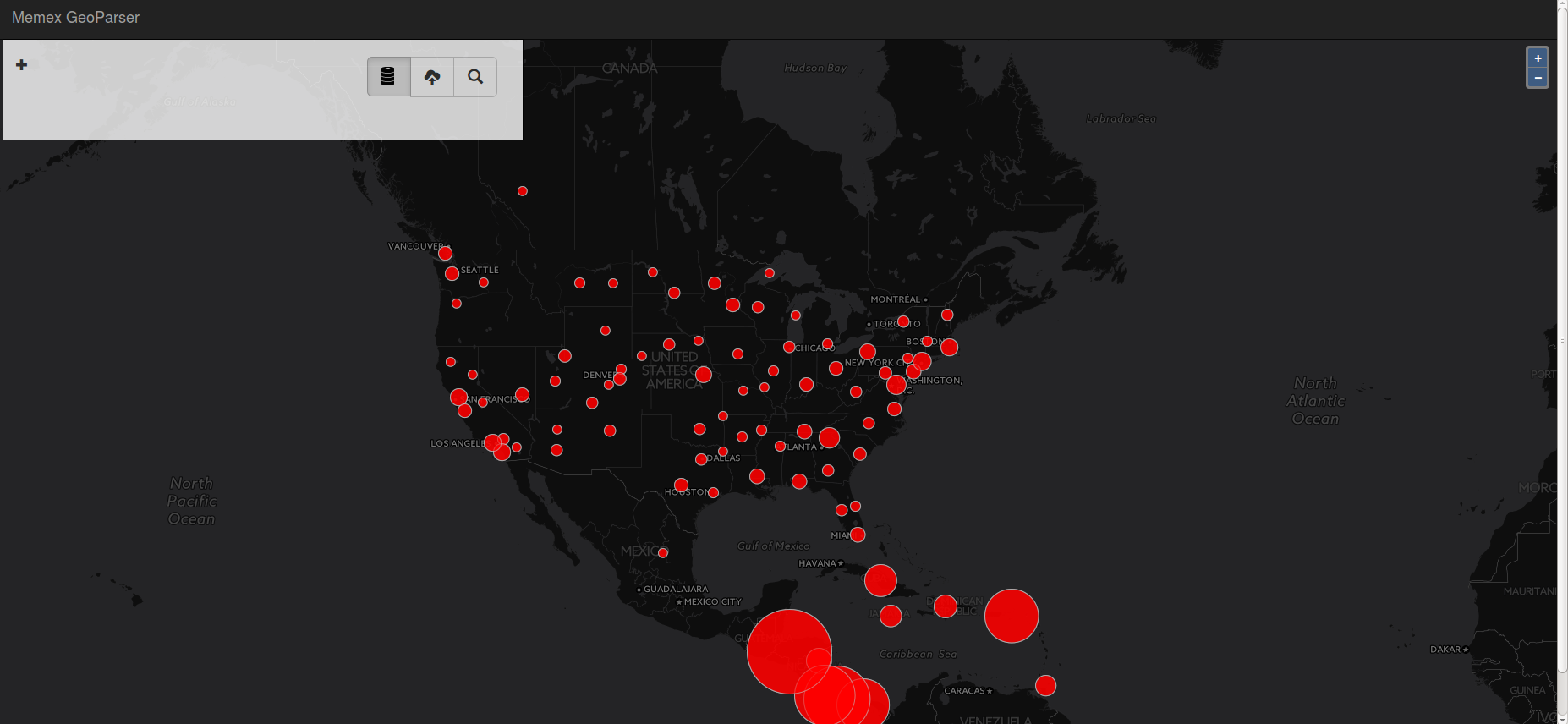
**DOI Content Handler**

We used yourls, which is a URL shortner library, to generate shortened URLs for each file. Once the proper setup for the yourls API was set up, it was quite easy to generate the shortened URLS. We read each file and sent the details to the yourls API which returned the shortened URL.

**Memex GeoParser**

Using the data from the GeoTopic parser, we indexed the data into solr. Then we ran the Memex GeoParser on the index and generated the location map. The screenshots of the location map that was generated are given below:





**D3 – Bar chart**

The bar chart gives the visual representation of metadata score for each file that is present in the solr index.