Abstract:

‘Crowd Based Data Collection Platform’ is infrastructure for citizen science projects. Project evolves around the idea getting data from required areas and also to save citizen’s time in collecting data. Citizen Science projects are developed for research purpose and citizens contribute in research by collecting and submitting it, so that further analysis can be done by scientists involved in project. Here Citizen plays main role in data collection. So, it is necessary to get relevant data from citizen’s which is useful and required by scientists for research purpose. If we get repeated and data which is of no use from citizen’s then that data is simply discarded, but in this citizen’s time and energy is wasted.

Also, in many projects data depends on area it is being collected. By this it means, Scientist needs data from different areas in a region. But, it may happen that they(scientists) get data from same area. These data is of no use to scientists and is simply discarded. Because of this aim of exploring all areas of particular region is not satisfied. So, here both scientists and citizens both are at loss, as citizens time is wasted by submitting the data from same area which is already explored and also scientists doesn’t get relevant information. In this project we came up with a solution to this problem. We developed an infrastructure for projects whose data depends on area it is being collected. Main idea of this project was to get data from unexplored areas. Basic idea of project is explained in introduction and motivation. Formal problem statement is mentioned in Problem statement section. Technical details are also included in this report. We also evaluated our results which are mentioned in evaluation section. There is also future scope to this project which we included in future work section.

Introduction and Motivation:

This project is developing an infrastructure for crowd sourcing projects. General idea of crowd sourcing projects is to collect data required for research purpose from citizens. Many projects require data from many regions, going to all regions and collecting data by all project members is nopossible, that’s why crowd sourcing citizens help to collect and provide it to scientists who will be using these data for research purpose. In this way citizens indirectly contribute to research. Involvement of citizens play a major role in success of crowd sourcing projects.

There were many issues involved in developing crowd sourcing project such as increasing participants, educating them about collecting correct data, etc. One such major problem is to get unnecessary data from citizens. This project is trying to solve this problem, by getting only necessary data.Many crowd sourcing projects data depends on area it is being collected. For example, data of “Marine Debris” project. In marine debris project citizens are collecting data from different lakes. In this it may happen that we may end up getting a data of specified area of particular lake site repeatedly instead of all areas of lake site being covered. This might also be the case that data being collected will not be useful for research. This project focuses on getting data of necessary area, instead of getting data from some areas repeatedly.

Main motivation for this project was to get data from unexplored areas. Many projects have tried solving problem of getting data from same area repeatedly. For example, ”Crowd-based Smart Parking” project directs user to unexplored area to get data from that area. But, for this users location is constantly being traced, and according to their current location they are directed to unexplored area. This will lead to security concerns, as users locations are being traced. Many projects like “Leafsnap” and “Lake observer” provide user with a location where data were collected by other users. However, this again will not cover unexplored areas.

Our project “Crowd-Based Data Collection Platform” ensures that data of required area is only collected, while keeping in mind user's security concern. By this project we will develop an infrastructure in which scientists will mark the area they want the data for and will alert the users that we need data for this area. This will allow to get data of unexplored area, instead of same area repeatedly. We also allow scientists to specify the stop limit for data, so that no extra data should be collected for that area. For this project we are considering air pollution data as use case. While this concept can be tested on various type of data.

Problem Statement:

For particular region scientists marks the area he is interested in getting data for. Also scientist specifies the number of data entries required for that marked area(Stop Limit). When user submits the data with his location information, then following things are done:

1. Check whether user's location falls in area marked by the scientists. If yes then only include in database, else show the user the areas the scientist is actually interested in.
2. If data being submitted crosses the stop limit for particular marked area, does not include that data.
3. Discard repeated data.

Technical Core:

We used google maps API to support maps related functionality. We have used marker to point to particular location, rectangle to display rectangle on map, drawing to draw on maps i.e. scientists marks the area on map and we also we got longitude and latitude by through google maps API. We used google maps controls through javascript in web pages. We allowed user to marks his position through marker. We Stored latitude and longitude of areas marked by scientists in MySQL database. Then we retrieved this marked areas for particular region from database to check whether user falls in area marked by the scientists using javascript.

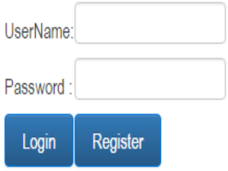
Database a major role in this project. We are storing scientists as well as users information in database. For example, area's location details. We are also storing data submitted by user in database. Here air pollution data. We are using JSP Servlets for business logic processing. In this html and javascript is used as front end, Servlets for logic processing and database for storing data. It is a web application project. In this we have used layered architecture to separate the concerns while developing the application and to incorporate changes easily.

We have included screen shots of our project which will give the clear picture of our project.

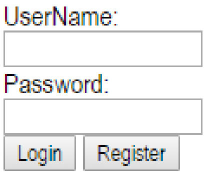
Home Screen:

Welcome to Data Collection Platform

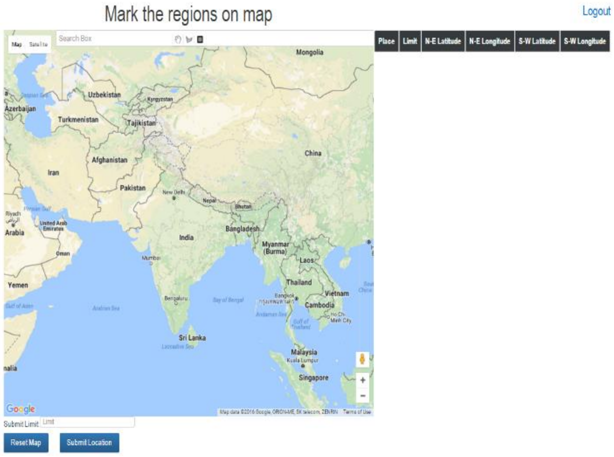
Scientist Login page:



Citizen Login Page:



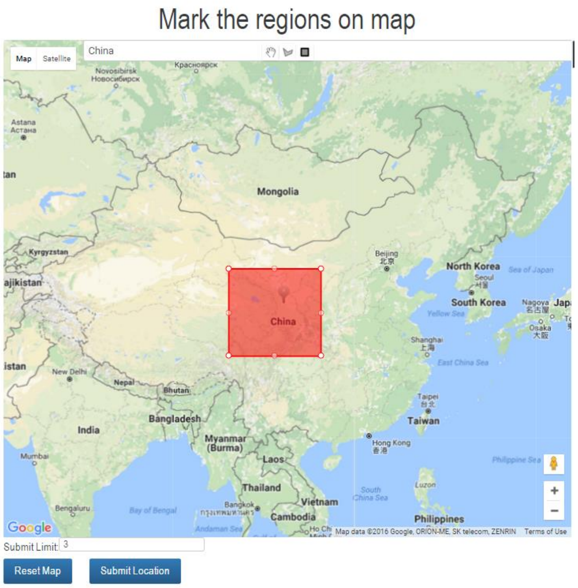
Scientist home page:



Mark the regions on map

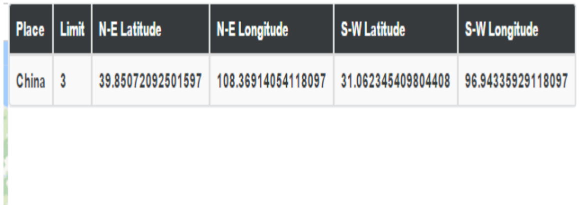
Place	Limit	N-E Latitude	N-E Longitude	S-W Latitude	S-W Longitude

Mark the region on the map, enter the stop limit in the text box:



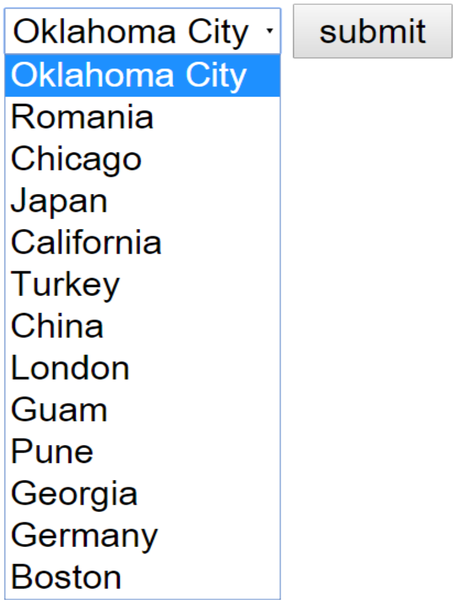
Submit Limit: 3

Details getting updated to the database:

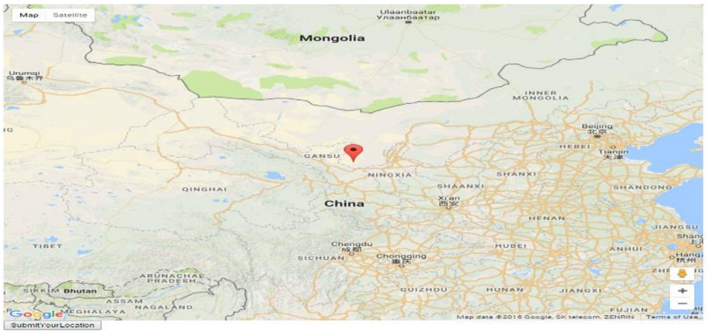


Place	Limit	N-E Latitude	N-E Longitude	S-W Latitude	S-W Longitude
China	3	39.85072092501597	108.36914054118097	31.062345409804408	96.94335929118097

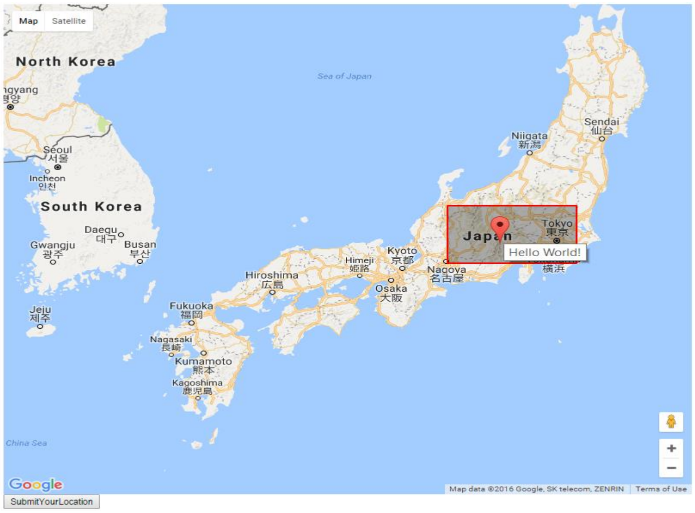
Select The Region From marked Areas:



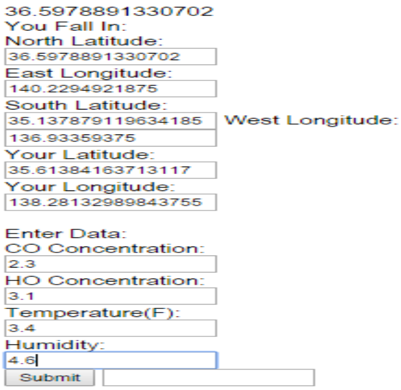
User's Location:



User's Location within marked area:



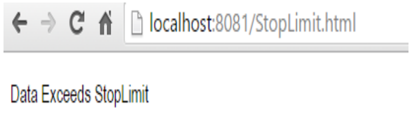
Form(Air Pollution Data):



36.5978891330702
You Fall In:
North Latitude:
36.5978891330702
East Longitude:
140.2294921875
South Latitude:
35.137879119634185 West Longitude:
136.93359375
Your Latitude:
35.61384163713117
Your Longitude:
138.28132989843755

Enter Data:
CO Concentration:
2.3
HO Concentration:
3.1
Temperature(F):
3.4
Humidity:
4.6

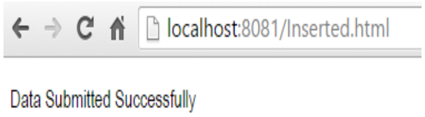
Data Exceeds Stop Limit:



localhost:8081/StopLimit.html

Data Exceeds StopLimit

Data Submitted Successfully:



localhost:8081/Inserted.html

Data Submitted Successfully

Evaluation

As Crowd Based Data Collection Platform is a Citizen Science project, we are following similar benchmark standards for evaluating our project. For the time being we are evaluating our project only in academic point of view and this may be changed later based on how we are going to extend or generalize our project. The goal of our project is ensure a productive participation by the citizens and make them contribute their intellectual or physical efforts in providing the data to the Research Scientist. For this purpose, we deploy an initial prototype and get feedback from the users and conduct a survey on usability of application. Based on this, the project can be improved further by improving the usability and user engagement.

Also, the project should be able to engage users from diverse locations and citizens should be able to enter data irrespective of their background. This means that anyone can become a citizen scientist. And the project should be developed in such a way that it preserves the interests of both the scientists and the citizen scientists. In addition to this everyone involved should make a sense out of what is going on with the project and how the data that the citizens are going to provide is used further by the scientists.

Based on these grounds, we are evaluating the project and the following are the categories under which we ensure that our project is as per the standards:

Usability*:* The user interface should be easy to use and as dynamic as possible. Also, we developed the application is such a way that the citizen scientist and research scientist can use the same application and we are taking care of separating the responsibilities. If a user clicks on Scientist option it redirects him/her to the Scientist module and one can login from there and continue further. On the other hand, if the user clicks on the Citizen option, he/she is redirected to the Citizen module. The separation of responsibilities is clear and there is no scope of confusion anywhere in the application.

Performance*:* One of the evaluation measures is performance and we will check on how the system is performing when multiple users are giving how input, output results and alerts are appearing in a timely manner.

Scalability: We are going to check to how much extent it is scalable, based on increasing the number of citizens and scientists involved with a particular project and how the platform as a whole is handling the load.

Fault tolerance: As part of evaluating fault tolerance, we are going to check how the system can regain information when website or database goes down.

In addition to these general standards, we evaluate our system using unit testing and integration testing. The following are the test cases that we checked to find how the system is behaving:

1. Check if the user is giving an input within the specified time frame by the scientist.
2. Check the stop limit constraint for the entries that are coming from different users. For example, if the already 4 entries are available for a location, we are going to discard the 5th entry and notify the user on this.
3. Check if the user is providing the data for a required location or not.
4. Check if a user is entering any duplicate data. If the new entry is identified to be a duplicate of already existing data, we are going to discard

that entry and notify the user that it is a duplicate.

5. Check if the scientist is able to mark the location as per his/her interest and if the locations are realistic.

6. Because this is an application that utilizes google maps, we are also testing if the locations or areas marked are similar to other applications that utilize google maps.

7. We are using the Air Quality Dataset format for evaluating the data in the forms that the citizen scientist enters.

Based on our evaluation and test results we conclude that our project is working as per other citizen science projects with improved usability along with the new ideas we added.

Conclusion

This application is a common data collection platform that can be generalized for collecting any kind of scientific data based on observations at user locations. Such an application minimizes the cost and efforts required by the scientists alone. Also, the diversity in terms of locations can be increased as people from every nook and corner can equally access the application.

There is a lot of scope for future work in the area and the application can be supported by multi-platforms like web and mobile also. Instead of making the user to mark the location, GPS tracking in one's mobile can be used to identify the user's location. Also, we can alert him/her once the user's location falls in the scientist's area of interest prompting them to enter any data if available.

Moreover, the citizens can subscribe to any events related to such projects or training and act in a proactive way to make the application and the interests work in a successful manner. Thus, the project can solve some of the crucial citizen science problems in an effective way.

References:

1. Crowdsourcing for On-street Smart Parking
Xiao Chen, Elizeu Santos-Neto, Matei Ripeanu
Department of Electrical and Computer Engineering, University of British Columbia
{xiaoc, elizeus,matei}@ece.ubc.ca

2. Creek Watch: Pairing Usefulness and Usability for Successful Citizen Science.