Final Project Report

Project Title:

Toxic release comparison for USA between years (2005 & 2006)

Team Information:

- **Members:**
  - Somasundaram Ardhanareeswaran
  - Alay Vora

Type of Project: LOGD

1. Introduction:

- The project focuses on using Semantic web approach to integrate two RDF datasets from Data.gov website using SPARQL queries. The integration happens with the help of a common predicate.
- The datasets being used for the project are been hosted on **Jena Fuseki Server** on an **Amazon EC2 instance.**
- Using the Google Visualization API's, the required results are rendered to the webpage using **JavaScript and HTML**.
- **Geo Map**, **Data Table** and **Column Chart** are the three **Google Visualizations** used for displaying the results.
- **Geo Map** shows the map of the United States which **change in Air Emissions between the years 2005 and 2006** for each State which can be seen while hovering the mouse to the respective State region.
- **On clicking the State** region in the Geo Map we can see a **bar chart** comparing the change of all **toxics released (Land, Underground and Air)** by the state into the environment.

2. Target Audience:

- This project can be useful to the Environmental Protection Agency of USA for studying the growth of toxic release across the years and to take corrective measures in this area.
- The categorized analysis can also be used to regulate environmental taxes in the respective state, where the toxic growth has dramatically increased.

3. Description of Data Sources:

- **Dataset 191: 2005 Toxics Release Inventory National data file of all US States and Territories**
    - The Toxics Release Inventory (TRI) is a publicly available EPA database that contains information on toxic chemical releases and waste management activities reported annually by certain industries as well as federal facilities.

- **Dataset 249: 2006 Toxics Release Inventory National data file of all US States and Territories**
    - The Toxics Release Inventory (TRI) is a publicly available EPA database that contains information on toxic chemical releases and waste management activities reported annually by certain industries as well as federal facilities.

- **Number of Triples:**
    - Dataset 191 – 9,973,192
    - Dataset 249 – 10,040,889

4. Data Integration:

- The two datasets are being integrated in the website using a common predicate from both datasets having predicate name as **'state'**.
- The SPARQL query used for getting the required results contained aggregate functions like SUM.
- Jena Fuseki Server supports all the required aggregate functions and thus we used Fuseki's custom SPARQL endpoint for hosting the two Datasets.
- The Fuseki server can be accessed in the following server:
    - http://52.207.219.191:3030/

5. Data Product Results:

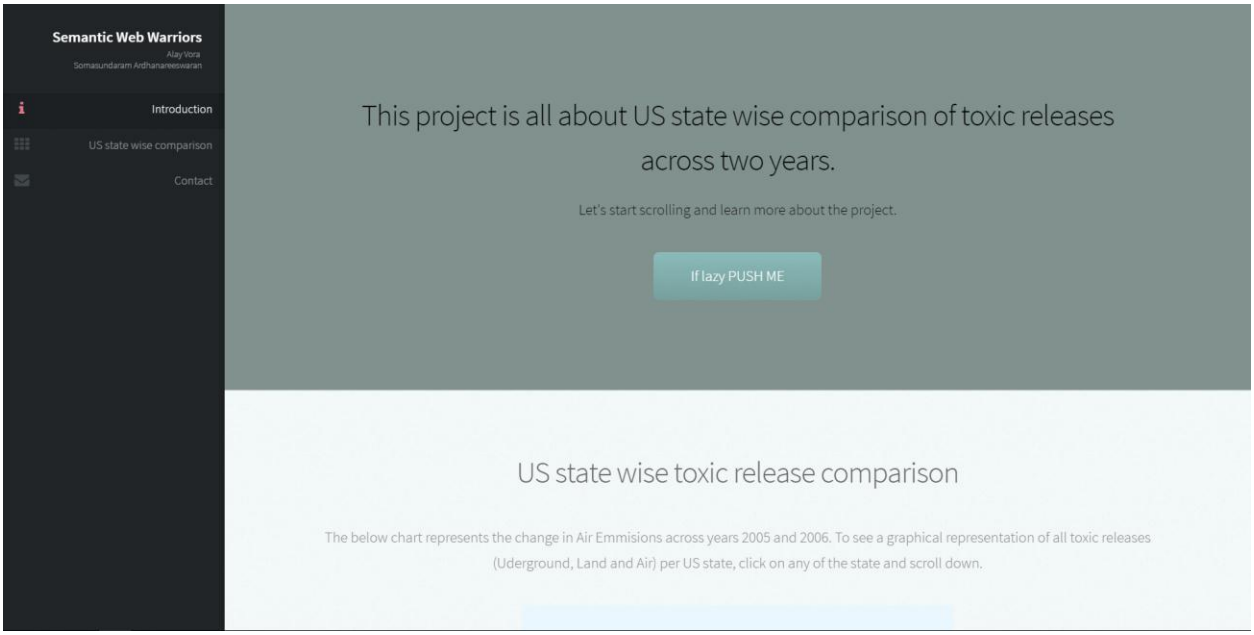

Figure 1: Homepage of the website describing the project

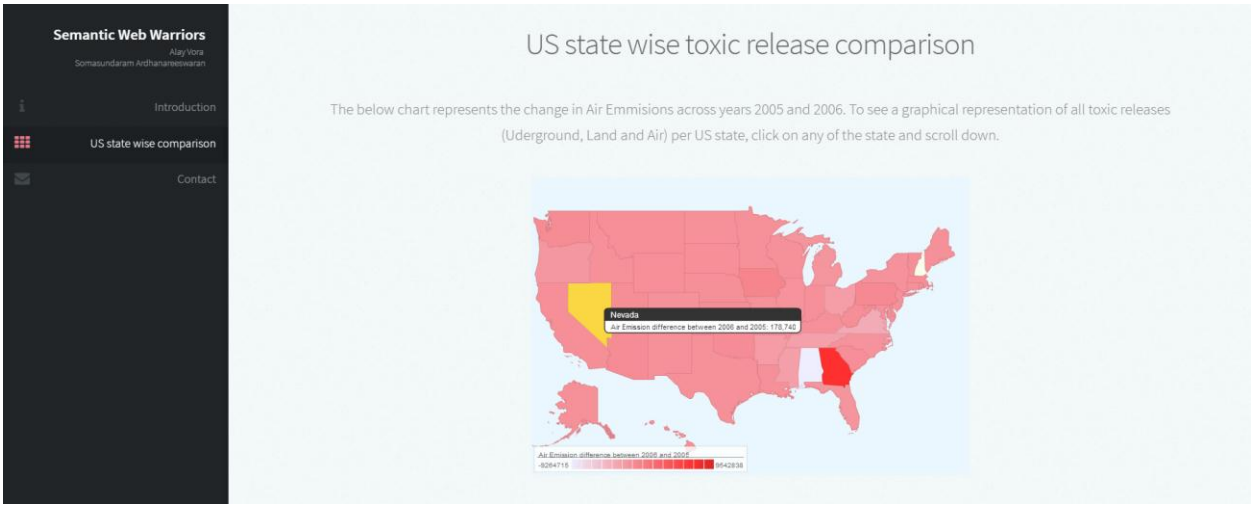

Figure 2: GeoMap representation of the change in Air Emissions between the years 2005 and 2006

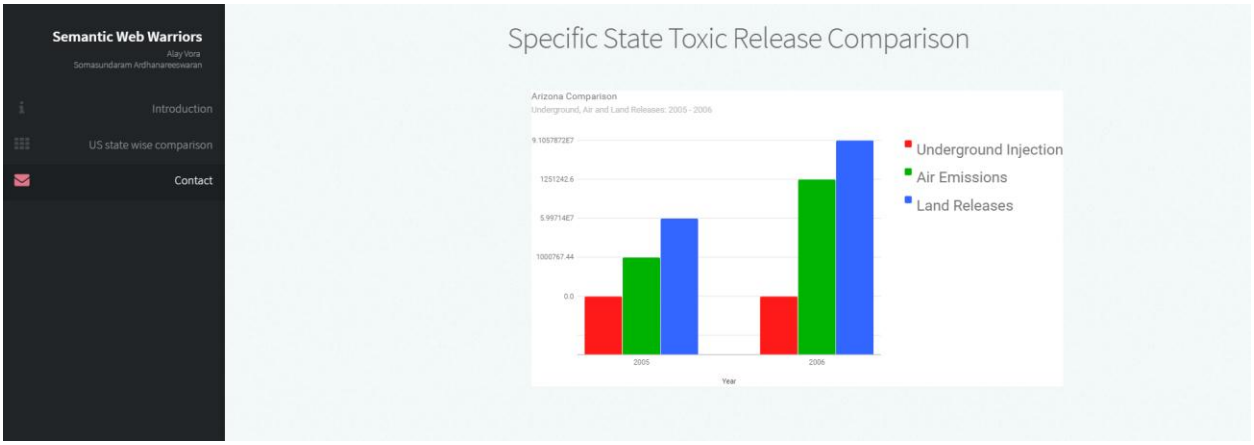

Figure 3: Underground, Air and Land Toxic release comparison for a specific state

6. Summary:

The project is to leverage the power of semantic web to show the world's biggest problem (pollution) in a graphical way. This representation clearly tells us which states in the US are environment friendly and which of them need attention to protect our planet.

One critical issue was that, Fuseki server is very memory intensive when we use such large datasets which led us to develop on a cloud instance with significantly higher memory.

There can be more representation of the same data, in terms of drilling down further inside a particular state, perhaps into the county or event into the city level. We can also have representations for each registered company and their recorded level of toxic release across the years and perform any corrective actions.