TITLE: DISEASE-PATHOGEN-LOCATION CORRELATION

TEAM INFORMATION:
TEAM NUMBER: #7
TEAM NAME: DISEASE PATROL
TEAM MEMBERS:
Aditi Venkatesh(AXV180047)
Anusha Pugazhendhi(AXP190006)

TYPE OF PROJECT: SIMPLE PROJECT

1. INTRODUCTION

Semantic web is a web of data that aims at making all data available on the World Wide Web machine-readable for use and sharing across multiple applications. In this project, we try to link data from multiple resources based on a few common patterns(attributes) we notice to provide some not-so-obvious insights. This is done with the help of a number of technologies including Resource Description Framework(RDF), Web Ontology Language(OWL), SPARQL and the Apache Jena Fuseki server.

Using State as a common attribute, using data from multiple disease records we try to figure out which state has more possibilities of pathogen-presence to disease-confirmation ratio so as to bring awareness to the need for fund allocation in the health-care sector for disease prevention. This is done by a simple ranking by conversion rates in ascending order of reporting locations(states).

2. TARGET AUDIENCE

- This project has been developed mainly for providing insights to the health-care sector as to which states are more susceptible to a disease wave and where more government funds need to be allocated to increase staffing and resources in hospitals, provide vaccinations etc.
- Since we found the conversion rates of pathogens to serious illness in each of these states, the information can be used to find out which of these states require more attention in terms of sanitation facilities and where the spread of a disease is more prevalent
- The state health department can use this information to possibly change sanitation laws and help maintain a healthy neighbourhood.
- More awareness programmes could be started to better facilitate healthy and clean habits in these states.

- We also find out how quick the diseases tend to spread in each of these states. For example, the data has information on cases recorded over regular intervals. This tells us how much time it takes for each of the diseases to fall under the dangerous category.

3. DESCRIPTION OF DATA SOURCES

- Tetanus to Trichinellosi

https://catalog.data.gov/dataset/nndss-table-1ii-tetanus-to-trichinellosis-bdb61

Publisher: data.cdc.gov Centre for disease control and prevention
RDF dump:
https://data.cdc.gov/api/views/2993-4v7d/rows.rdf?accessType=DOWNLOAD
Number of triples: 12604
The dataset describes for the year 2019, number of Tetanus cases in the 50 U.S. states. It also describes number of cases for the current week that the data has been recorded in, number of cases for the previous 52 weeks and the cumulative number of cases in 2019

- Spotted fever rickettsiosis, Confirmed to Smallpox

https://catalog.data.gov/dataset/nndss-table-1gg-spotted-fever-rickettsiosis-confirmed-to-smallpox-e25d3
Publisher: data.cdc.gov Centre for disease control and prevention
RDF dump:
https://data.cdc.gov/api/views/rihk-iawc/rows.rdf?accessType=DOWNLOAD
Number of triples: 12604
The dataset describes for the year 2019, number of Smallpox cases in the 50 U.S. states. It also describes the number confirmed cases for the current week, number of cases confirmed for the previous 52 weeks, number of probably cases in the past 52 weeks, cumulative cases in 2018, its corresponding flags and other columns that have not been made use of

- Cholera to Coccidioidomycosis

https://catalog.data.gov/dataset/nndss-table-1h-cholera-to-coccidioidomycosis-43b90
Publisher: data.cdc.gov Centre for disease control and prevention
RDF dump:
https://data.cdc.gov/api/views/wycc-ffia/rows.rdf?accessType=DOWNLOAD
Number of triples: 12604
The dataset describes for the year 2019, number of cholera cases in the 50 U.S. states. It also describes the number confirmed cases for the current week and its flag and the

number of cases confirmed for the previous 52 weeks and its flag and the cumulative cases in 2018-2019

- Foodbourne to Botulism

https://catalog.data.gov/dataset/nndss-table-1e-botulism-foodborne-to-botulism-other-c466c
Publisher: data.cdc.gov Centre for disease control and prevention
RDF dump:
https://data.cdc.gov/api/views/4t6w-ibvk/rows.rdf?accessType=DOWNLOAD
Number of triples: 12604
The dataset describes for the year 2019, number of botulism cases in the 50 U.S. states.It also describes the number confirmed cases for the current week and its flag and the number of cases confirmed for the previous 52 weeks and its flag, the cumulative cases in 2018-2019 and the same attributes for infants as well.

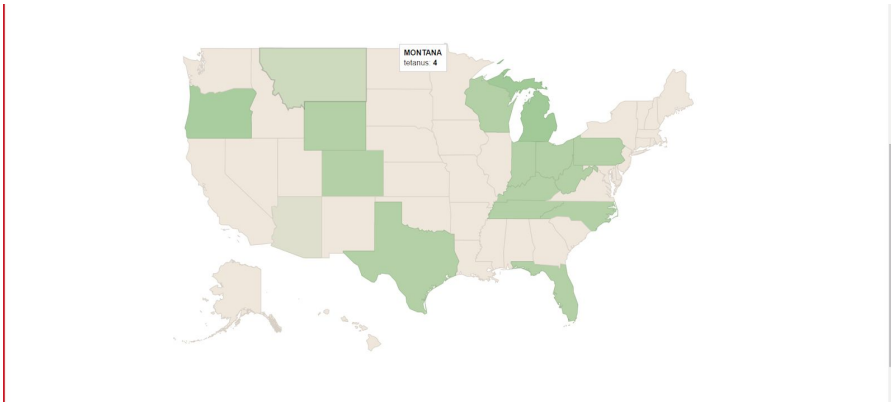
4. DATA INTEGRATION:

- All four datasets have a common attribute with URI <http://data.cdc.gov/resource/_2993-4b7d/reporting_area> that describes the location (state) at which the case has been identified. This has been used to mash up all four datasets to compare maximum probability of occurrence of these diseases, grouped by state.This was done using the metadata obtained by analyzing the datasets which helped in identifying the different parameters, and also how the data is mapped onto RDF based representation.

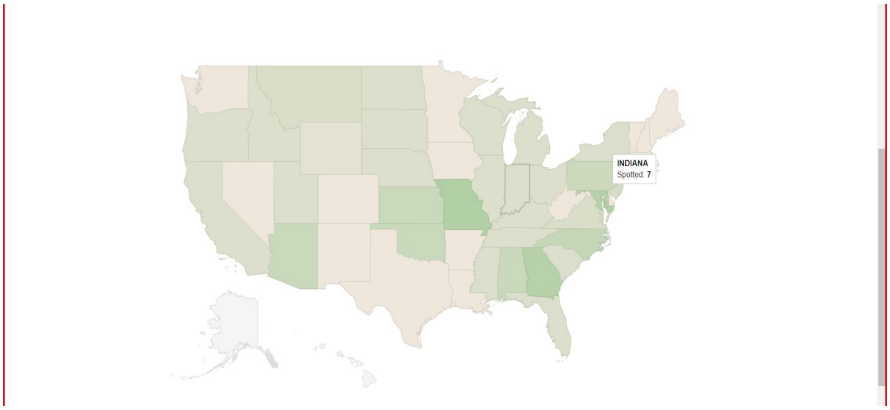
5. DATA PRODUCT RESULTS

User is provided with the HTML link to the main webpage where they can see the pie-charts visualised. SPARQL queries are written to order the number of cases of the four diseases state-wise for which the data is accessed using the local Apache FUSEKI server endpoint.

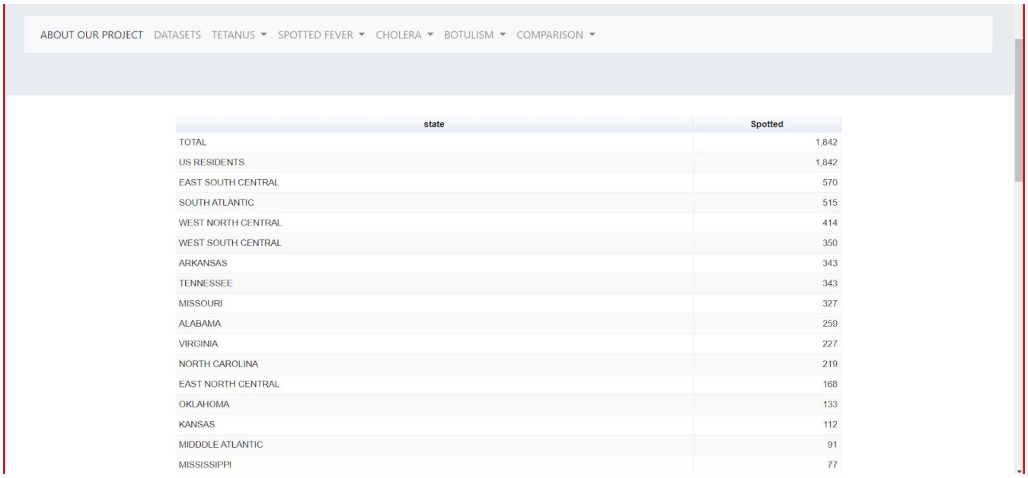
State-wise number of probable cases for Tetanus



State-wise number of probable cases for Spotted fever

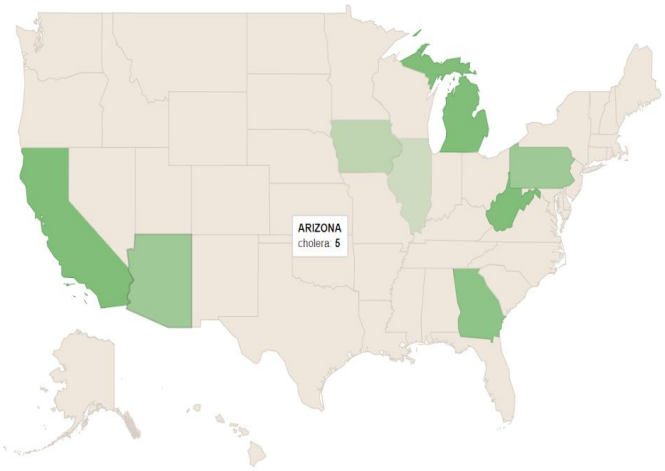


Probable cases spotted fever rickettsiosis in desc order of cases

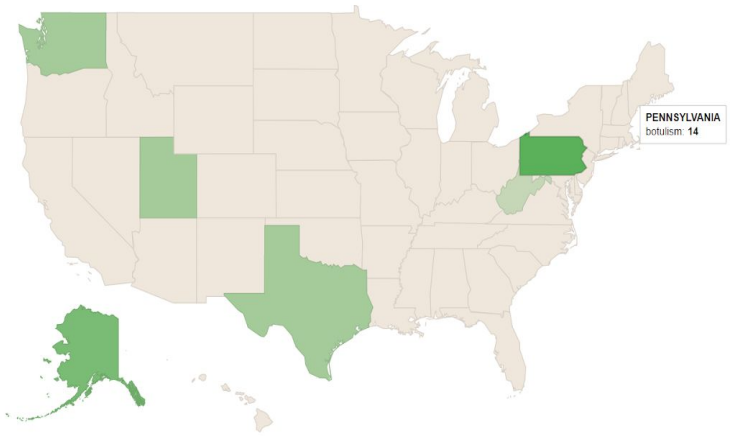


state	Spotted
TOTAL	1,842
US RESIDENTS	1,842
EAST SOUTH CENTRAL	570
SOUTH ATLANTIC	515
WEST NORTH CENTRAL	414
WEST SOUTH CENTRAL	350
ARKANSAS	343
TENNESSEE	343
MISSOURI	327
ALABAMA	259
VIRGINIA	227
NORTH CAROLINA	219
EAST NORTH CENTRAL	168
OKLAHOMA	133
KANSAS	112
MIDDDLE ATLANTIC	91
MISSISSIPPI	77

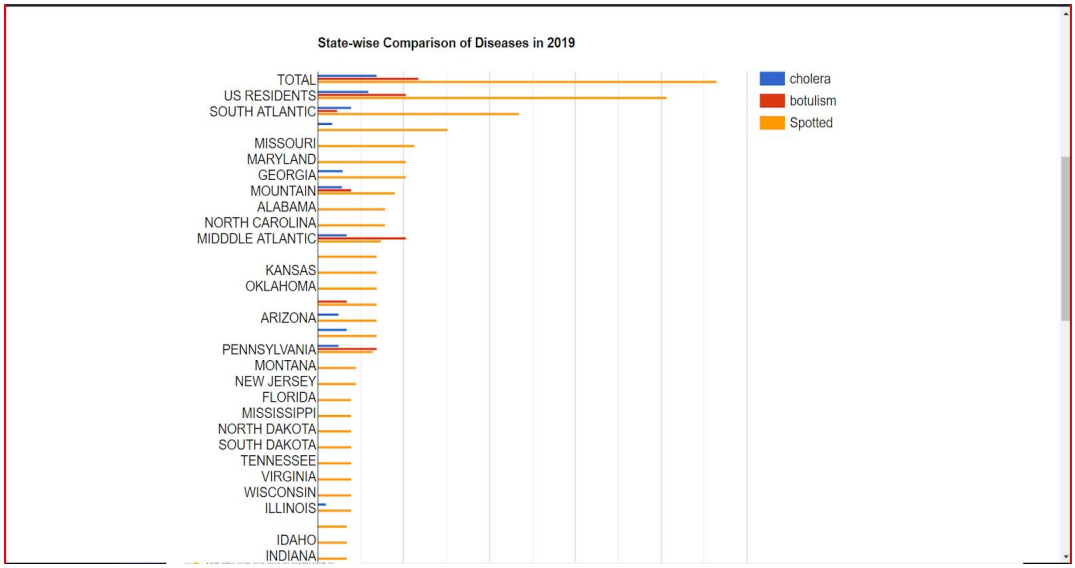
State-wise number of probable cases for Cholera fever



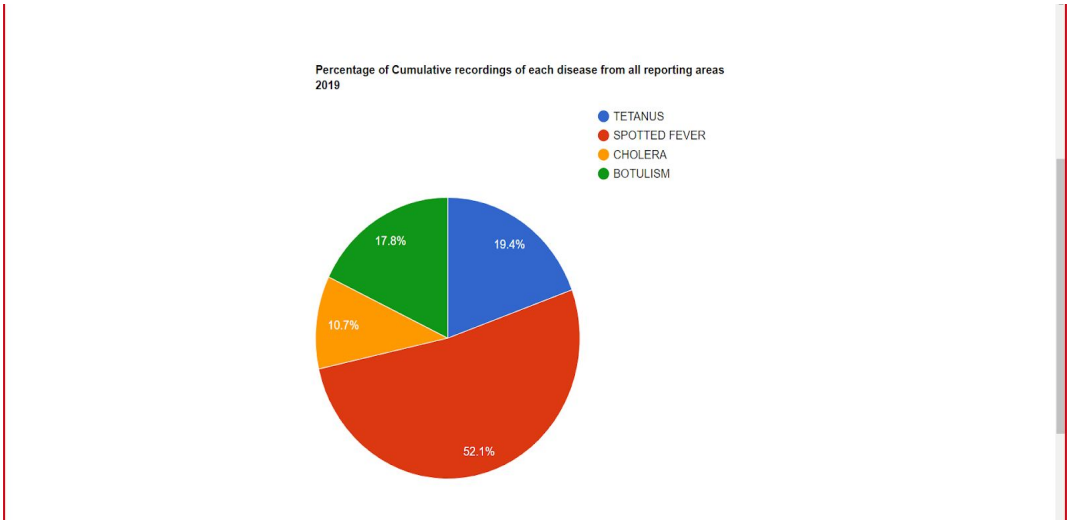
State-wise number of probable cases for Botulism fever



State-wise comparison of diseases



Percentage of diseases in all reporting areas



Insights from the project

- State-wise number of reports for all diseases (tetanus, spotted fever, cholera and botulism were found using google visualization API
- Maximum and Minimum no. of cases were found from each of the disease maps. The following table shows the states with the maximum number of cases of each of the 4 diseases:

Disease	State	No. of Cases
Tetanus	Michigan	9
Spotted Fever	Missouri	23
Cholera	Michigan,California	7
Botulism	Pennsylvania	9

We see that Michigan happens to have the most number of cases of both Tetanus and Cholera. From the visualizations we can also infer that there are some areas which do not have any of these diseases. For example Connecticut, Louisiana, Delaware have 0 recorded cases for each of these diseases.

While Arkansas has the maximum no. of probably cases for spotted Fever, It did not have any confirmed cases for the same.

- Overall comparison of 4 diseases from all reporting areas was shown using a pie chart. We see that spotted fever is a more prevalent disease among the 4 and cholera is rare

- Conversion ratio for a few selected areas:

State	Cholera conversion ratio	Botulism conversion ratio	Spotter fever conversion ratio
Arizona	0.0029	0.0	0.667
Michigan	0.4375	0.0	0.5
Missouri	0.0	0.0	0.07

The conversion ratio gives the ratio of the number of confirmed cases to probable cases. This talks about the chances of getting a disease in case they have the symptoms for it. The figures also show that botulism is almost obsolete now. Good job on that!

6. SUMMARY

OUTSTANDING ISSUES

- The datasets were large and it took a long time to process query results and even upload them
- Integrating datasets for diseases from different sources. We had to go through a lot of datasets to find those with matching attributes
- Ambiguity in data (missing values). There were a lot of values that were unavailable and it was difficult to get accurate results
- Problem accessing local Fuseki endpoint

LESSONS LEARNED:

- Better SPARQL querying techniques
- Ability to interpret RDF data
- Visualisation techniques