Final Report

**Interrelation between Federal Student Aid and Adult Literacy Rate**

# 

# Group Information

Group Name: Krishna Arjun

Type of Project: Custom Project

Project Name: **Interrelation between Federal Student Aid and Adult Literacy Rate**

Team Members: 2

## Details of team members

1. Shraddha Bang - SXB180041
2. Sankalp Bhandari - SSB170006

Table of Contents

[Group Information 2](#_Toc24561865)

[Details of team members 2](#_Toc24561866)

[Report 3](#_Toc24561867)

[1. Introduction 3](#_Toc24561868)

[2. Target Audience 3](#_Toc24561869)

[3. Description of data sources 3](#_Toc24561870)

[4. Data Integration 4](#_Toc24561871)

[5. Data Product Results 5](#_Toc24561872)

[6. Custom project Justification 11](#_Toc24561873)

[Summary 12](#_Toc24561874)

# Introduction

This project is about relation extraction from unstructured business data using semantic web technologies. It aims to derive unstructured data from the web and process content in the data to find entities and relationships. These entities and relationships, modelled by a semantic ontology, organizes the content in the data to a structured format. The results are then displayed on to a website, where the user can see the relation between different entities and obtain linked information about those entities.

This report discusses the target audience for this project. Then it explains the data and discusses about the custom ontologies used to describe the structured data. After that, some focus is laid on the data integration and its value and then some results are displayed.

The report finally justifies the custom project choice and concludes with a brief summary.

## Target Audience

This project can be relevant/useful to the following:-

1. This information will be useful for government officials to identify the areas which directly affected by lack of funds and hence, work on improving the condition.
2. This information may also be used by parents/students to identify developed areas which will ensure a quality education with better facilities

## Description of data sources

Data sets used:

* Datasets: Campus Based Program Data for Federal Student Aid

Provides annual recipient, disbursement and federal award information for the Campus-Based Programs by postsecondary school.

<https://catalog.data.gov/dataset/campus-based-program-data-for-federal-student-aid>

* Datasets: 2003 National Assessment of Adult Literacy

This data by NAAL is a cross-sectional assessment that collected information about English literacy among American adults age 16 and older.

<https://catalog.data.gov/dataset/2003-national-assessment-of-adult-literacy>.

The following ontology describes the structure in which data will be described in the results:-

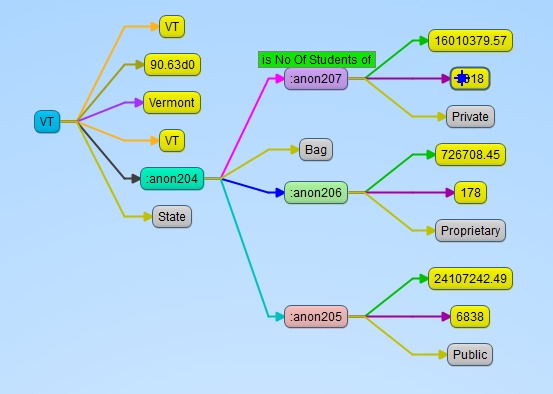


Figure 1: Custom Ontology Description

## Data Integration

Data is integrated in the following steps:-

1. Firstly, we downloaded the data from the above sources.
2. Then we found the useful features that may affect the literacy rate of a state.
3. Then this data is stored in csv files
4. We wrote custom program to create RDF ontology from the features in csv files
5. We uploaded these RDF files to Fuseki in order to create SPARQL endpoint
6. Then from our website, we query the SPARQL endpoint to get the data related to the charts which we want to show to provide relation between different features.
7. This integration from end-to-end provides the experience of transformation of unstructured data scattered over the web into a structured, linked format that the user can make sense of and derive meaningful information from.

## Data Product Results

Following was the expected results of this project:-

1. Getting an insight as state funding affects the literacy rate of states based on their school type.
2. We can also derive which school types are lacking funding and we can improve the funding for them.
3. The following charts will give government officials a single glance at data to identify problems and they can take related measures to overcome these problems.

These results are interesting because scope of this project can help government officials make better informed decision to distribute funding based on extracted relationships and structures data visualizations. Even casual users might benefit by gather general knowledge from this data about state literacy.

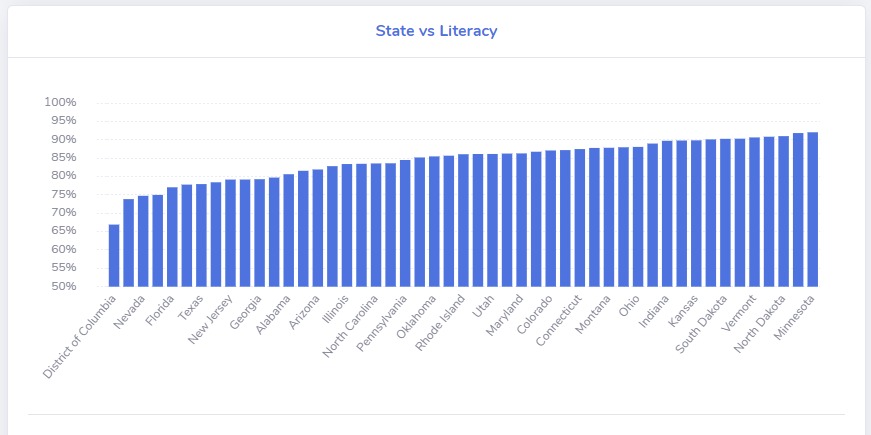


Figure 2: State vs Literacy chart

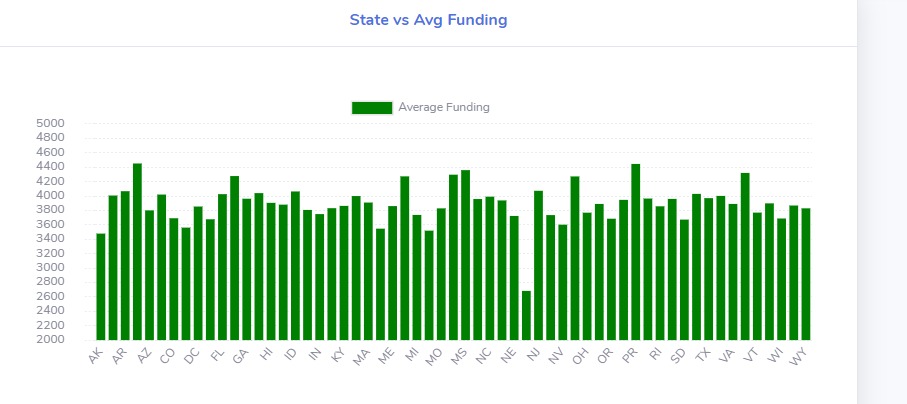


Figure 3: State vs Avg Funding chart

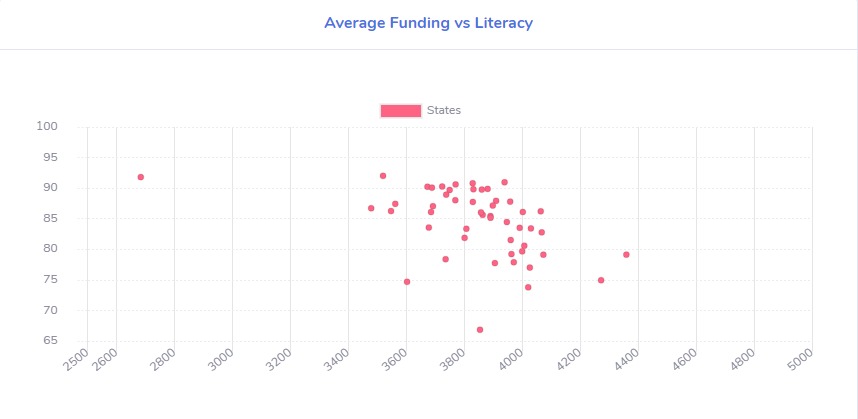


Figure 4: Avg Funding vs Literacy chart

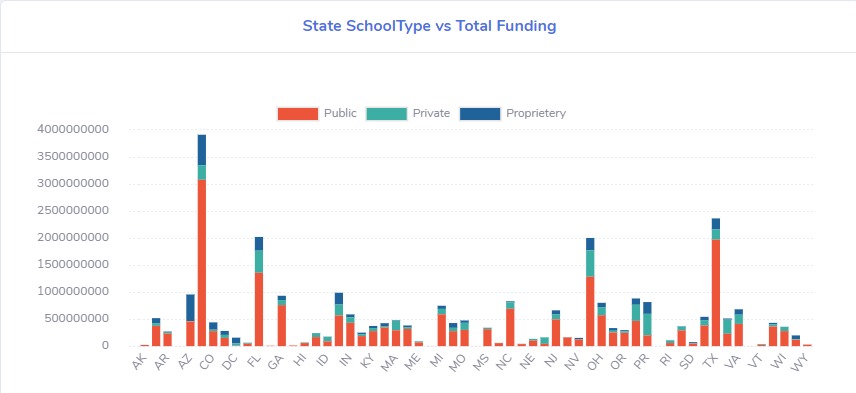


Figure 5: State School Type vs Total Funding chart

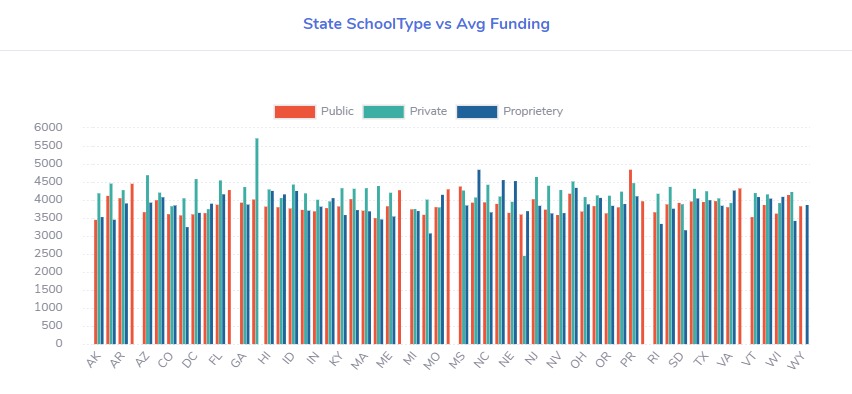


Figure 6: State School Type vs Avg Funding chart

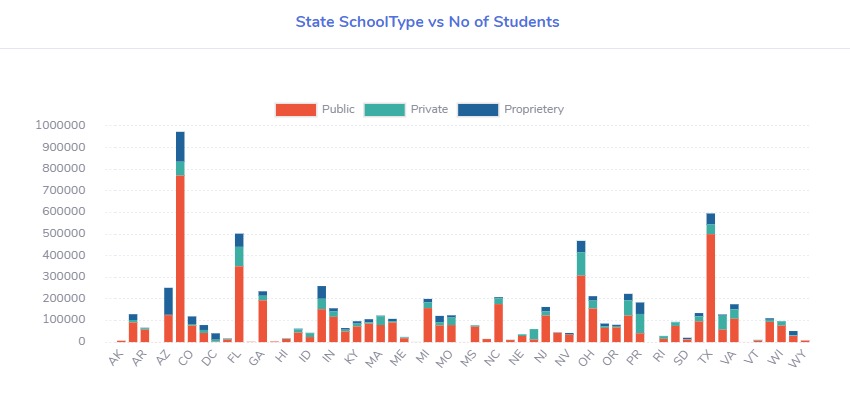


Figure 7: State School Type vs No of Students chart

The above graph stored as RDF format looks as follows: -

**StateLiteracy.rdf**

<rdf:Description rdf:about="http://www.krishna-arjun.org/#NV">

<rdf:type rdf:resource="http://www.krishna-arjun.org/class#state"/>

<ns1:literacyrate rdf:datatype="http://www.w3.org/2001/XMLSchema#double">74.74</ns1:literacyrate>

<ns1:abbr>NV</ns1:abbr>

</rdf:Description>

**StateMapping.rdf**

<rdf:Description rdf:about="http://www.krishna-arjun.org/#AS">

<rdf:type rdf:resource="http://www.krishna-arjun.org/class#state"/>

<ns1:abbr>AS</ns1:abbr>

<ns1:name>American Samoa</ns1:name>

</rdf:Description>

**StateSchoolType.rdf**

?xml version="1.0" encoding="UTF-8"?>

<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:ns1="http://www.krishna-arjun.org/">

<rdf:Description rdf:about="http://www.krishna-arjun.org/#AK">

<rdf:type rdf:resource="http://www.krishna-arjun.org/class#state" />

<ns1:schooltype>

<rdf:Bag>

<rdf:li>

<ns1:Public>

<ns1:NoOfStudents>6272</ns1:NoOfStudents>

<ns1:Disbursement>21638956.92</ns1:Disbursement>

</ns1:Public>

</rdf:li>

<rdf:li>

<ns1:Proprietary>

<ns1:NoOfStudents>374</ns1:NoOfStudents>

<ns1:Disbursement>1318978</ns1:Disbursement>

</ns1:Proprietary>

</rdf:li>

<rdf:li>

<ns1:Private>

<ns1:NoOfStudents>228</ns1:NoOfStudents>

<ns1:Disbursement>954422.68</ns1:Disbursement>

</ns1:Private>

</rdf:li>

</rdf:Bag>

</ns1:schooltype>

</rdf:Description>

On the website, we have also shown the queries used to fetch the data for the charts. They are shown as below.



Figure 8: State School Type vs No of Students chart

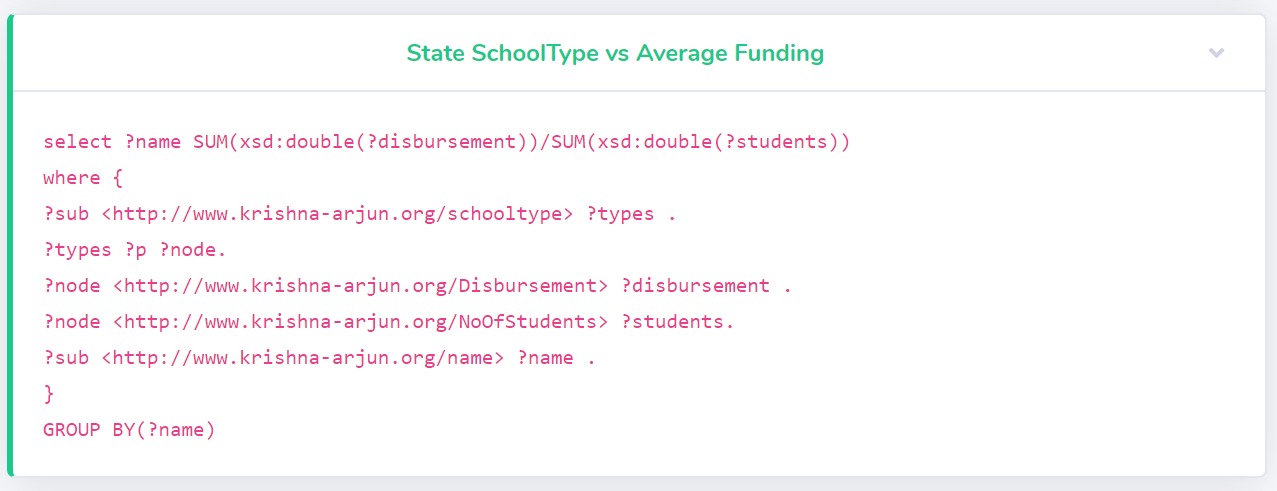
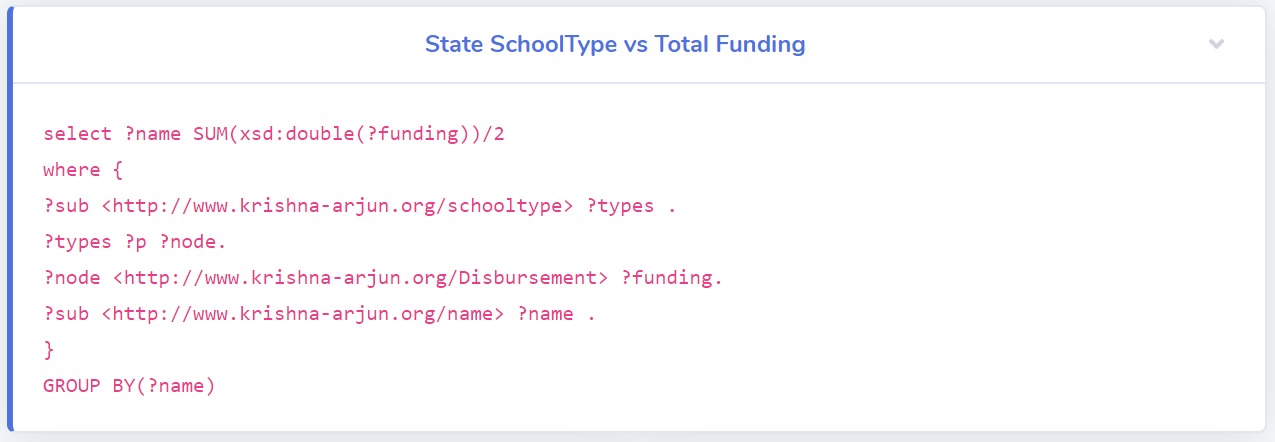
Figure 9: State School Type vs Avg Funding 

Figure 10: State School Type vs Total Funding

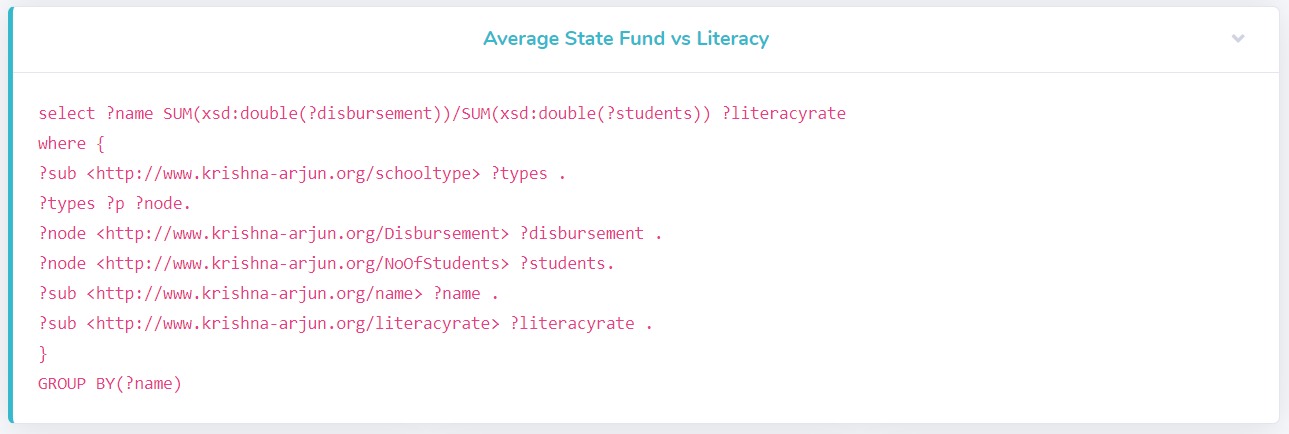


Figure 11: Avg Funding vs Literacy

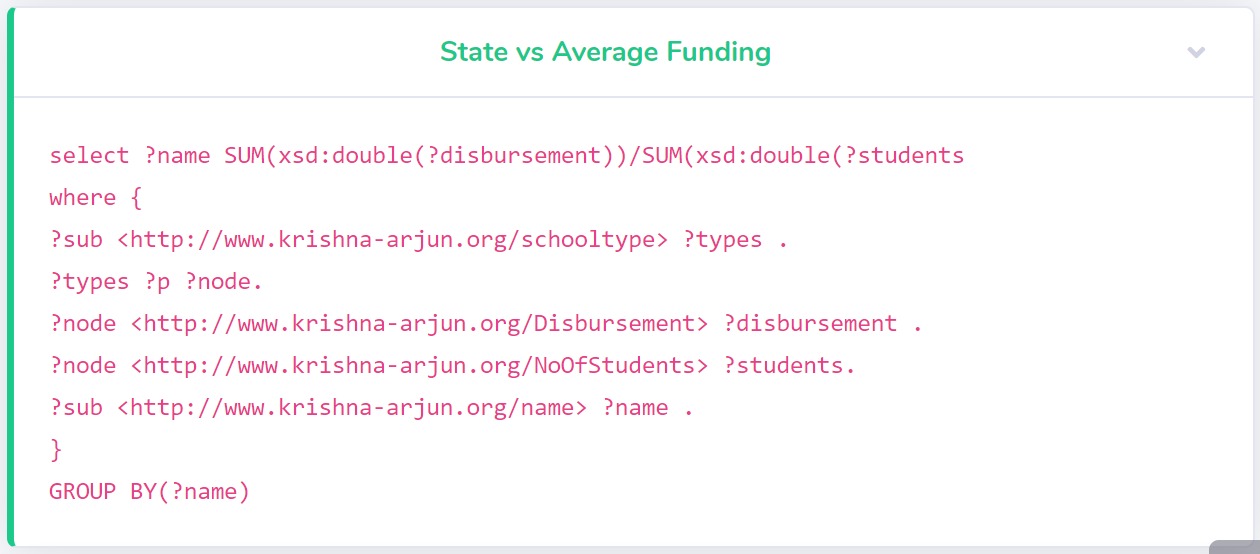


Figure 12: State vs Avg Funding

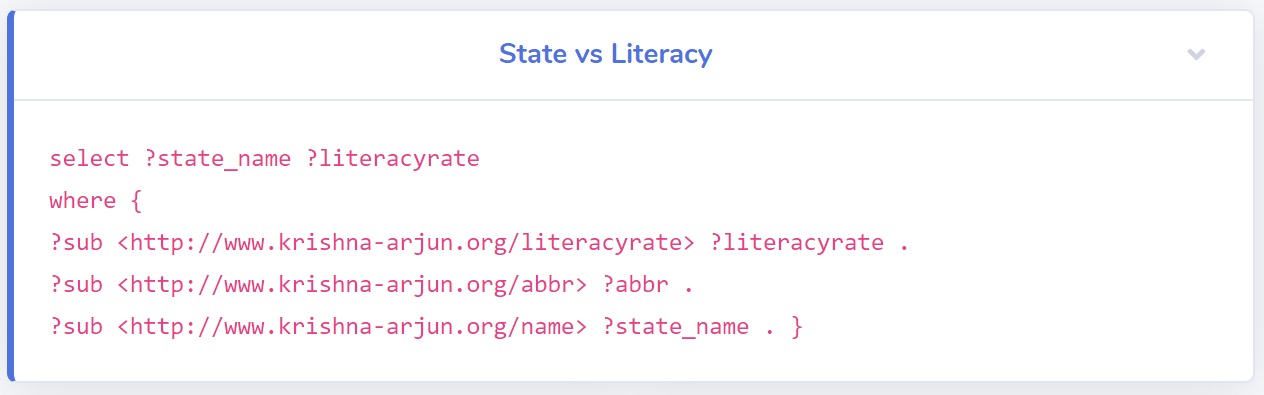


Figure 13: State vs Literacy chart

## Custom project Justification

This project differs from an LOGD project in the following ways:-

1. Different source of data – We extracted the data which was relevant to us and did not use the data as it is.
2. SPARQL Endpoint: We have created our own sparql endpoint using the data which we query from our website in realtime.
3. Custom Ontology – A custom ontology was defined to model and structure the data and the results displayed were structured using this model.

# Summary

As discussed, this project successfully extracts relations in unstructured data using Semantic Web technologies. As shown in the figures, these relations in the data and successfully displayed on a website for the user to derive meaningful information from.

The report also described the relevance of these results, the target audience for whom these result may prove to be relevant and the justification for this custom project. With the use of technologies as discussed in the report and the accuracy of the results obtained, this project can very well be useful for the audience it aims to target.