Project 3: Semantic Data Retrieval

Team 3: Liam Nguyen, Loc Hinh, Phuc Dang, Suchitra Chinnamail and Varun L

Link: https://github.com/liam-nguyen/CECS-571/tree/master/project_3

Previously

- We built a hospital ontology with 3 datasets
- All from data.gov
- All in 2018
- All in CSV

Hospital General Information

- Description: a list of all hospitals that have been registered with Medicare including:
  - ID
  - Name
  - Contact Information (address, phone number, state, etc…)
  - Services
  - Type
  - Overall rating

Medicare Hospital Spending by Claim Type

- The data presented on Hospital Compare provide price-standardized, non-risk-adjusted values for hospital spending by claim type.
  - Facility ID
  - Hospital average Medicare spending
  - State average Medicare spending
  - National average Medicare spending

Timely and Effective Care

- Description: this data set includes provider-level data for measures of cataract surgery outcome, colonoscopy follow-up, heart attack care, emergency department care, preventive care, pregnancy and delivery care, and cancer care.
  - Facility ID
  - Score

Semantic Data Retrieval

- Original decision (abandoned):
  - Spring Boot (backend)
  - React (frontend)
  - RDF4J formerly Sesame (SPARQL query)
- Instead, use Apache Fuseki for everything.

Apache Jena Fuseki

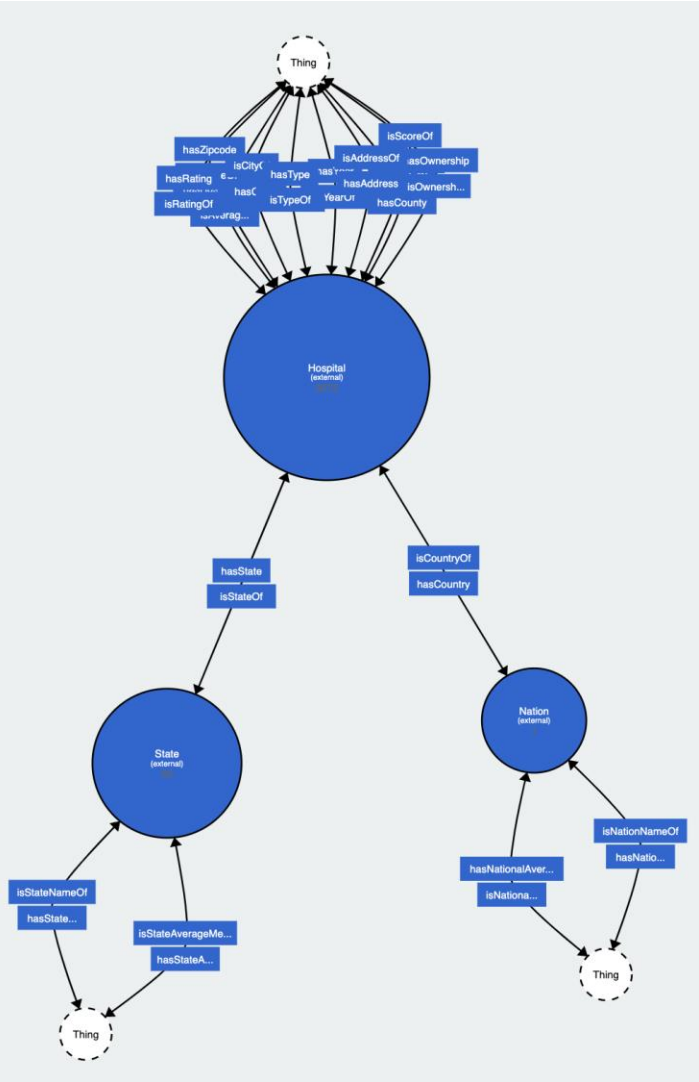
- It's a SPARQL server as a Java web application
- Provide SPARQL 1.1 protocol
- Simple to configure and launch
- Link: https://jena.apache.org/documentation/fuseki2/
- Just download and launch fuseki-server script.
- If you get permission denied (for mac/linux): chmod +x fuseki-server

Our Ontology

Class and Property Pair

	class	property	
1	ds:Hospital	ds:hasYear	Hospital
2	ds:Hospital	ds:hasZipcode	
3	ds:Hospital	ds:hasScore	
4	ds:Hospital	ds:hasFacilityID	
5	ds:Hospital	ds:hasHospitalAverageMedicareSpending	
6	ds:Hospital	ds:hasCity	
7	ds:Hospital	ds:hasType	
8	ds:Hospital	ds:hasState	
9	ds:Hospital	ds:hasRating	
10	ds:Hospital	ds:hasOwnership	
11	ds:Hospital	ds:hasCountry	
12	ds:Hospital	ds:hasPhoneNumber	
13	ds:Hospital	ds:hasFacilityName	
14	ds:Hospital	ds:hasCounty	
15	ds:Hospital	ds:hasEmergencyService	
16	ds:Hospital	ds:hasAddress	
17	ds:State	ds:hasStateName	State
18	ds:State	ds:hasStateAverageMedicareSpending	
19	ds:Nation	ds:hasNationalAverageSpending	Nation
20	ds:Nation	ds:hasNationName	

Visualization



Question 1 - Suchitra

In 2018, which states have average Medicare spending above the national average?

```
PREFIX  xsd:  <http://www.w3.org/2001/XMLSchema#>
PREFIX  rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX  rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX  owl:  <http://www.w3.org/2002/07/owl#>
PREFIX  ds:   <https://data.medicare.gov/d/nrth-mfg3#>

SELECT  ?stateName ?averageMedicareSpending ?NationAverageMedicareSpending
WHERE
  { ?state  ds:hasStateName       ?stateName ;
            ds:hasStateAverageMedicareSpending  ?averageMedicareSpending
    { SELECT  (AVG(?medicareSpending) AS ?NationAverageMedicareSpending)
      WHERE
        { ?state  ds:hasStateAverageMedicareSpending  ?medicareSpending
          FILTER ( ?medicareSpending != 0 )
        }
    }
    FILTER ( ?averageMedicareSpending > ?NationAverageMedicareSpending )
  }
```

Question 2 - Loc

Which proprietary hospital provides emergency psychiatric services? (Loc)

```
PREFIX  xsd:  <http://www.w3.org/2001/XMLSchema#>
PREFIX  rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX  rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX  owl:  <http://www.w3.org/2002/07/owl#>
PREFIX  ds:   <https://data.medicare.gov/d/nrth-mfg3#>

SELECT  ?subject ?facility_name ?facility_address ?emergency_service ?service_type ?ownership
WHERE
  { ?subject  ds:hasFacilityName    ?facility_name ;
              ds:hasAddress         ?facility_address ;
              ds:hasEmergencyService  "true" ;
              ds:hasEmergencyService  ?emergency_service ;
              ds:hasType            "Psychiatric" ;
              ds:hasType            ?service_type ;
              ds:hasOwnership       "Proprietary" ;
              ds:hasOwnership       ?ownership
  }
```

Question 3 - Phuc

Given hospitals with higher spending than the national average, what's the percentage of those that have higher score than national average score?

```
PREFIX  xsd:  <http://www.w3.org/2001/XMLSchema#>
PREFIX  rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX  rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX  owl:  <http://www.w3.org/2002/07/owl#>
PREFIX  ds:   <https://data.medicare.gov/d/nrth-mfg3#>

SELECT  ?percentage_nation
WHERE
  { { SELECT  (COUNT(?_id) AS ?req_hopitals)
      WHERE
        { ?subject  ds:hasFacilityID      ?_id ;
                    ds:hasHospitalAverageMedicareSpending  ?hospital_spending ;
                    ds:hasCountry         ?country .
          ?country  ds:hasNationalAverageSpending  ?nation_spending .
          ?subject  ds:hasScore           ?score
          { SELECT  (round(AVG(?score)) AS ?avg)
            WHERE
              { ?hospital  ds:hasScore  ?score }
          }
          FILTER ( ( ?hospital_spending > ?nation_spending ) && ( ?score > ?avg ) )
        }
    }
    { SELECT  (COUNT(?id) AS ?total_hospital)
      WHERE
        { ?subject  ds:hasFacilityID      ?id ;
                    ds:hasHospitalAverageMedicareSpending  ?hospital_spending ;
                    ds:hasCountry         ?country .
          ?country  ds:hasNationalAverageSpending  ?nation_spending
          FILTER ( ?hospital_spending > ?nation_spending )
        }
    }
    BIND(( ( ?req_hopitals / ?total_hospital ) * 100 ) AS ?percentage_nation)
  }
```

Question 4 - Varun

Which hospital has the best ratio of spending to score? (Varun)

```
PREFIX  xsd:  <http://www.w3.org/2001/XMLSchema#>
PREFIX  rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX  rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX  owl:  <http://www.w3.org/2002/07/owl#>
PREFIX  ds:   <https://data.medicare.gov/d/nrth-mfg3#>

SELECT  ?name ?score ?spending ?stateSpending (( ?spending / ?score ) AS ?ratio)
WHERE
  { ?hospital  ds:hasState          ?state .
    ?state     ds:hasStateName       ?stateName .
    ?hospital  ds:hasFacilityName    ?name ;
               ds:hasHospitalAverageMedicareSpending  ?spending ;
               ds:hasScore           ?score .
    ?state     ds:hasStateAverageMedicareSpending  ?stateSpending
  }
ORDER BY DESC(?ratio)
LIMIT   1
```

Question 5 - Liam

What is the most efficient state?

- Most efficient state = state that has the highest percentage of efficient hospitals of all hospitals.
- Efficient hospital = hospital that has higher score than state's average score and lower spending than state's spending?

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX ds: <https://data.medicare.gov/d/nrth-mfg3#>

SELECT ?stateName (ROUND(?efficient_count * 100 / ?total) AS ?percent) ?efficient_count ?total
WHERE {
  {
    # This is to find # efficent hospitals in each state
    SELECT ?stateName (COUNT(?efficient) AS ?efficient_count)
    WHERE {
      ?efficient ds:hasHospitalAverageMedicareSpending ?hSpending.
      ?efficient ds:hasScore ?hScore.
      ?efficient ds:hasState ?state.
      ?state ds:hasStateName ?stateName.
      ?state ds:hasStateAverageMedicareSpending ?sSpending.

      # This is to find out the state average score
      {
        SELECT (AVG(?innerScore) AS ?stateAvgScore)
        WHERE {
            ?hospital ds:hasState ?state.
            ?state ds:hasStateName ?stateName.
            ?hospital ds:hasScore ?innerScore.
         }
      }

      # Filter out by comparison
      FILTER(?hSpending < ?sSpending && ?hScore > ?stateAvgScore)
    }
    GROUP BY ?stateName
  }

  # This block is to find the total # hospitals in each state
  {
    SELECT ?stateName (COUNT(?hospital) AS ?total)
    WHERE {
```

Technical Difficulties

- Protégé unable to load ontology due to large file size or plugins --> Apache Fuseki
- What to do with differences in number of hospitals between dataset?
  - Reason to keep all: answer more questions
  - Reason to filter out: simpler SPARQL query (don't have to filter out values in aggregate query)