# Assignment 4

I will cover DCAT during the week of 6/22. Check out the Project Topics section for more details. Essentially DCAT will allow us to find all related data across multiple data.gov/ federal data sites. Try to find related data across databases (see data.gov)  to answer  slightly complicated questions. For example, you will have to use census data to find ethnic backgrounds and ages in various US regions and data from HealthData.gov on mortality and morbidity rates, or data on educational institutions (via data.gov) . Then, pose some question such as: Is health or education coverage worse in some regions and relate it to data from the US Census Bureau. This is just a suggestion. The idea is to use DCAT front-end to access data from two government data sets, as per your passion and background, and relate them using simple data analytics. A simple data analytic is to find the average for the entire US and define arbitrarily half that of that as the threshold  to define under-coverage, or one--and-a-half that value as the threshold to define over-coverage. The primary goal here is to integrate data across two data sets, using the DCAT interface. Any simple type of data analytics is fine. Do screen captures, and upload the code, screen captures, and a short paper to Github.

Submit it in a separate folder at your Github Repo entitled as "Assignment 4"

## Project

**DCAT - Data Catalog Vocabulary**

DCAT ( [http](http://www.w3.org/TR/vocab-dcat/" \t "_blank)://www.w3.org/TR/vocab-dcat/ ) has been approved as the  metadata schema for federal data catalogs - the front end to make it easy to find a database (which may still be in .csv etc format, and read out in JSON). Transition was supposed to be completed by 2/15. Here is a good link with multiple resources:<https://project-open-data.cio.gov/v1.1/schema/> . Use Python or Java semantic web framework (both to be covered before the end of July) to build a querying engine. For Python, it would use FuXi; for Java, it would use Jena (which is much more comprehensive). This is of strong interest to the federal government. They would like to see the process become easy and prevalent. Publishable and has job potential.

[www.digitalgov.gov/resources/how-to-get-your-open-data-on-data-gov/](https://exchange.fau.edu/owa/redir.aspx?C=D8SahN-eakW7iFUy5YQDmnBnN-OHkNIIgTw4Q3V204G0zfhD3jYJrabhK8JeTG852MA3ZOYhqRE.&URL=http%3a%2f%2fwww.digitalgov.gov%2fresources%2fhow-to-get-your-open-data-on-data-gov%2f)

There's also a post that goes over the updates to the v1.1 schema and includes a webinar for government agencies at

<http://www.digitalgov.gov/2014/12/04/data-govs-data-pipeline-explained/>

For the following, please copy and paste. Control and Click does not work for the following (I did not have time to change the linking):

The underlying platform used by Data.gov is CKAN ([http://ckan.org/](https://exchange.fau.edu/owa/redir.aspx?C=D8SahN-eakW7iFUy5YQDmnBnN-OHkNIIgTw4Q3V204G0zfhD3jYJrabhK8JeTG852MA3ZOYhqRE.&URL=http%3a%2f%2fckan.org%2f)) which a number of other governments use and others may be doing more on the linked data side than we are. The EU is also working on an update to their profile of DCAT which might be of interest - [https://joinup.ec.europa.eu/node/137964/](https://exchange.fau.edu/owa/redir.aspx?C=D8SahN-eakW7iFUy5YQDmnBnN-OHkNIIgTw4Q3V204G0zfhD3jYJrabhK8JeTG852MA3ZOYhqRE.&URL=https%3a%2f%2fjoinup.ec.europa.eu%2fnode%2f137964%2f)

CKAN also has some built in support for DCAT RDF XML. If you add ".rdf" to the end of any dataset URL on [catalog.data.gov](https://exchange.fau.edu/owa/redir.aspx?C=D8SahN-eakW7iFUy5YQDmnBnN-OHkNIIgTw4Q3V204G0zfhD3jYJrabhK8JeTG852MA3ZOYhqRE.&URL=http%3a%2f%2fcatalog.data.gov) you can get a DCAT RDF XML version, e.g. [https://catalog.data.gov/dataset/data-gov-visitor-metrics](https://exchange.fau.edu/owa/redir.aspx?C=D8SahN-eakW7iFUy5YQDmnBnN-OHkNIIgTw4Q3V204G0zfhD3jYJrabhK8JeTG852MA3ZOYhqRE.&URL=https%3a%2f%2fcatalog.data.gov%2fdataset%2fdata-gov-visitor-metrics) ->[https://catalog.data.gov/dataset/data-gov-visitor-metrics.rdf](https://exchange.fau.edu/owa/redir.aspx?C=D8SahN-eakW7iFUy5YQDmnBnN-OHkNIIgTw4Q3V204G0zfhD3jYJrabhK8JeTG852MA3ZOYhqRE.&URL=https%3a%2f%2fcatalog.data.gov%2fdataset%2fdata-gov-visitor-metrics.rdf)

## A data catalog conforms to DCAT if:

1. It is organized into datasets and distributions.
2. An RDF description of the catalog itself and its datasets and distributions is available (but the choice of RDF syntaxes, access protocols, and access policies is not mandated by this specification).
3. The contents of all metadata fields that are held in the catalog, and that contain data about the catalog itself and its dataset and distributions, are included in this RDF description, expressed using the appropriate classes and properties from DCAT, except where no such class or property exists.
4. All classes and properties defined in DCAT are used in a way consistent with the semantics declared in this specification.
5. DCAT-compliant catalogs may include additional non-DCAT metadata fields and additional RDF data in the catalog's RDF description.

# Assignment 4:

## Experiment:

-----------

Comparing 2 Datasets

- 1st Dataset from Data.gov includes "Estimates of the population by age, sex, race, and Hispanic origin for the nation, states, and counties" ;

http://catalog.data.gov/dataset/population-estimates

- 2nd Dataset from HealthCare.gov includes "Mortality Rate By Region" ;

http://catalog.data.gov/dataset/infant-mortality-by-race-ethnicity-2004-2013

My goal was to combine these two datasets and compare the infant mortality rate between the different races (White, Black, American Indian, Asian, Hispanic, Pacific Islander, And Two or More Races).

One dataset showed the total deaths there were for all infants across the different races. The second dataset showed the population estimate across the United States for all ages and races.

I provided a method to narrow the population estimate to show only the infants in the United States so I could see the total number of infants in the US for each race.

This number was essentially the "Total Number of Births" in the US for each race.

## Data Analysis Performed:

-------------------------

To find the infant mortality rate for each race, I had to first find the infant population for each race across the US to use as a base denominator.

Secondly, I had to find the number of total infant deaths for each race. The number of total deaths divided by the total number of births (infant population) returned the mortality rate for each race.

## Results:

-----------

After parsing the two RDF files within python and combining the 2 datasets, I have discovered that in 2013 the rates for infant mortalities for each race were as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Year** | **Race\_Ethnicity** | **Total Deaths** | **Total\_LiveBirths\_Denominator** | **Rate** |
| 2013 | American Indian | 9 | 1794 | 5 |
| 2013 | Asian | 170 | 68117 | 2.5 |
| 2013 | Black | 275 | 25837 | 10.6 |
| 2013 | Hispanic | 1212 | 238200 | 5.1 |
| 2013 | Not Stated or Unknown | 10 | 9385 | 1.1 |
| 2013 | Pacific Islander/Hawaiian | 6 | 1994 | 3 |
| 2013 | Two or more races | 133 | 11169 | 11.9 |
| 2013 | White/Other | 533 | 137896 | 3.9 |
| 2013 | Total | 2348 | 494392 | 4.7 |

Conclusion:

-----------

Infants with **two or more races** had the highest rate of mortalities with a rate of 11.9 deaths for every 1000 births.

**Black infants** came in second with a rate of 10.6 deaths for every 1000 births.

**Asian** infants had the lowest mortality rate at 2.5 deaths for every 1000 births.

