

**INM713 - *Semantic Web Technologies and Knowledge Graphs*  
Coursework Project**

Zacharias Detorakis

[zacharias.detorakis@city.ac.uk](mailto:zacharias.detorakis@city.ac.uk)

**Abstract**

This report aims at describing the steps taken and assumptions made as part of the coursework for the Semantic Web Technologies & Knowledge Graphs module. The goal of the coursework is to create a basic ontology that can fit the data from a csv file and generate the triples into a Knowledge Graph (KG). Following that there are tasks for reasoning, aligning the created ontology with an existing pizza ontology and querying the results triples to get insights from the underlying data. As a final task the OWL2Vec tool is used to create embeddings and used to create clusters of the ontology entities.

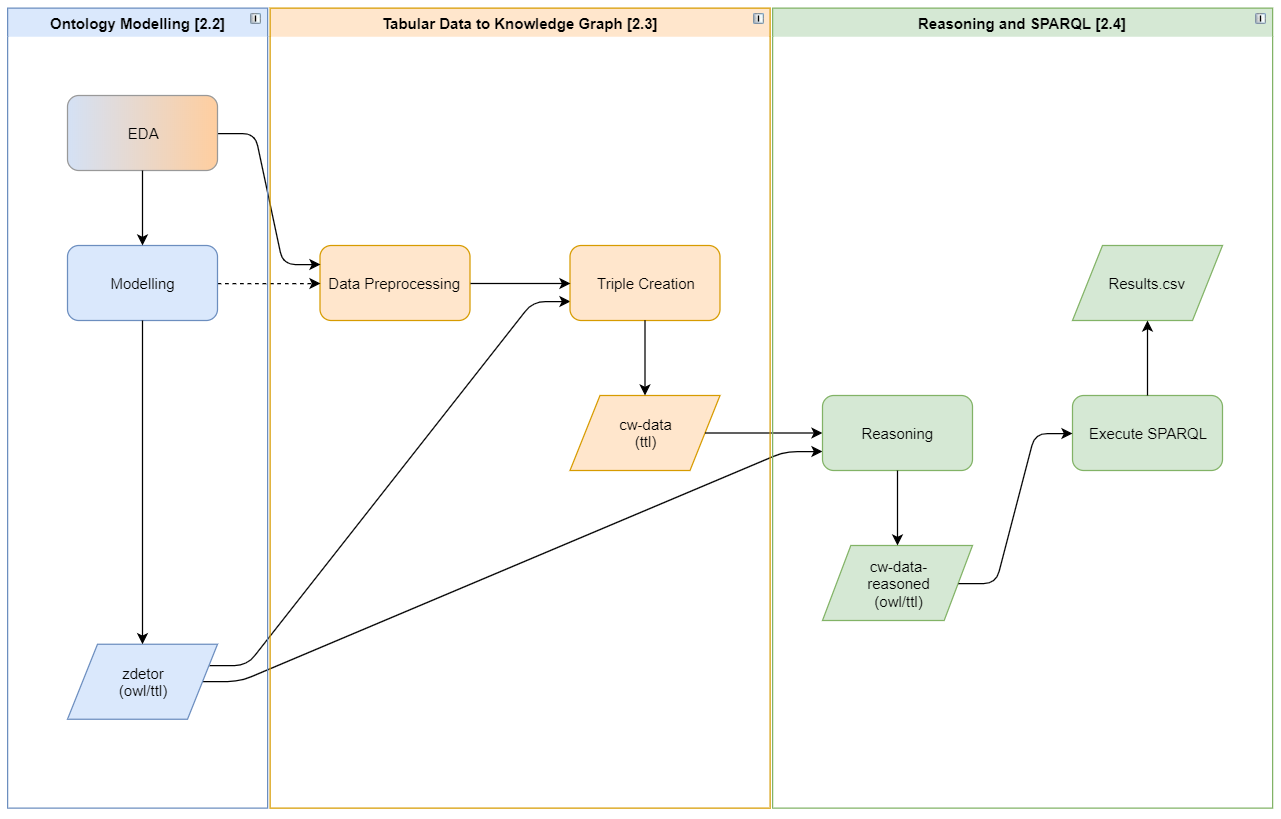


Figure 1. Process of modelling and converting tabular data to a KG

# Introduction

The following diagram [Figure 1] illustrates the steps taken for the first 3 tasks of the coursework (i.e. to create an ontology based of the give data, load the data to a KG and extract useful insights).

# Ontology Modelling [2.2]

## Exploratory Data Analysis (EDA)

Prior to jumping straight into modelling the first step was to understand the nature of the data we are trying to follow, and data discrepancies and most importantly what type of information that data holds.

Following is a high-level signature of the dataset

|  | **Description** |
| --- | --- |
| ***Number of Rows:*** | 3,510 |
| ***Grain:*** | The grain of the data in the csv file is at the menu item level. This means that each row represents a menu item (a pizza in the majority of the cases). There are also columns pertaining to the restaurant that serves each of those menu items. This information has been de-normalised and appears with the same set of values across all items served at each restaurant. |
| ***Columns:*** |  |
| name | *Number of unique restaurant names*: 933  Exploratory data analysis has shown that the name doesn’t uniquely identify the restaurant since there are a few sites with multiple addresses (e.g. post codes) and each address has it’s own set of menu items.  For instance, the site name Bertucci's appears with 4 different addresses and although three of them appear to be of the same franchise (i.e. having similar menu items with the same description and price), one of the three is missing the Margherita Pizza. Moreover the 4th one, the one in Illinois seems to be a completely different restaurant since it only has one menu item that is different from the menu items of the other 3 Bertucci’s restaurants.  The above observations have highlighted the need to combine the name with at least one more column before we can create a new instance of a site otherwise the data would incorrectly assume that because the name is the same all 4 Bertucci's sites are actually the same site.  Also, it has rendered the assumption that all restaurants sharing the same name are part of the same franchise/chain incorrect. Therefore, there is not enough information to safely create a hierarchy of restaurants belonging to the same chain (e.g. Bertucci’s, PapaJohn’s, Applebee’s) |
| address | The address line of the address the restaurant is in. This field, same as the name, is always populated so is a good candidate to use for a URI describing the restaurant |
| city | The city the restaurant is in |
| country | The country the restaurant is in. This column is always populated with the value ‘US’ which means all restaurants in the csv file are located in the United States |
| postcode | The address post code  Nulls: 65 |
| state | The state the restaurant is in. This column appears to have data of all sorts, so we decided to investigate the data a bit further. To avoid noise from the menu items, we focused on the section of the data that described the address and found the following:  *Number of unique addresses*: 989  *Unique values for state*: 281  *State values not 2 characters long*: 29.97%  *State values exactly 2 characters long*: 70.03%  Given that all the addresses are in the US and our prior knowledge that there are 52 states in the US, the 281 unique values in the state column imply that the column is sometimes used to store other type of information.  Given that the majority of the addresses (i.e.70%) have the char-2 state code, we will try to clean the remaining 30% during the creation of the URIs |
| categories | This column has some textual description of the category the restaurant belongs to. |
| menu item | The name of the menu item. This field is always populated so we can used it for the URI creation. However, the values are repeated across multiple restaurants (similar to the restaurant name) so we need to be careful when creating the individuals |
| item value | Nulls: 562  Not all items have prices. However, we are making the decision not to filter out any of the items missing a price as they may be offered for free. Moreover, we do not want to miss any of the information from the spreadsheet |
| currency | Nulls: 559  Same as the null values for the item values which goes to show that whenever we have a price, we also have a currency. It this was not the case we may choose to infer the currency from the country of the restaurant, but we will not do that for this case |
| item description | Nulls: 1984  A short description of the menu item. In most cases this column has a lot of useful details on the pizza ingredients. |

Table . Dataset signature

## Modelling

Having done an initial analysis of the data the next step is to create entities for the identified classes, object and data properties. As mentioned earlier the key information of the given dataset are menu items (pizzas) sold at restaurants across the US. Therefore, the key classes that will be created are ‘Restaurant’ and ‘Pizza’.

The following diagram [Figure 2] presents an individual of the *Restaurant* class “Bertucci’s” and an individual of the *Pizza* class “Margherita” as well as all other entities consider for the ontology.

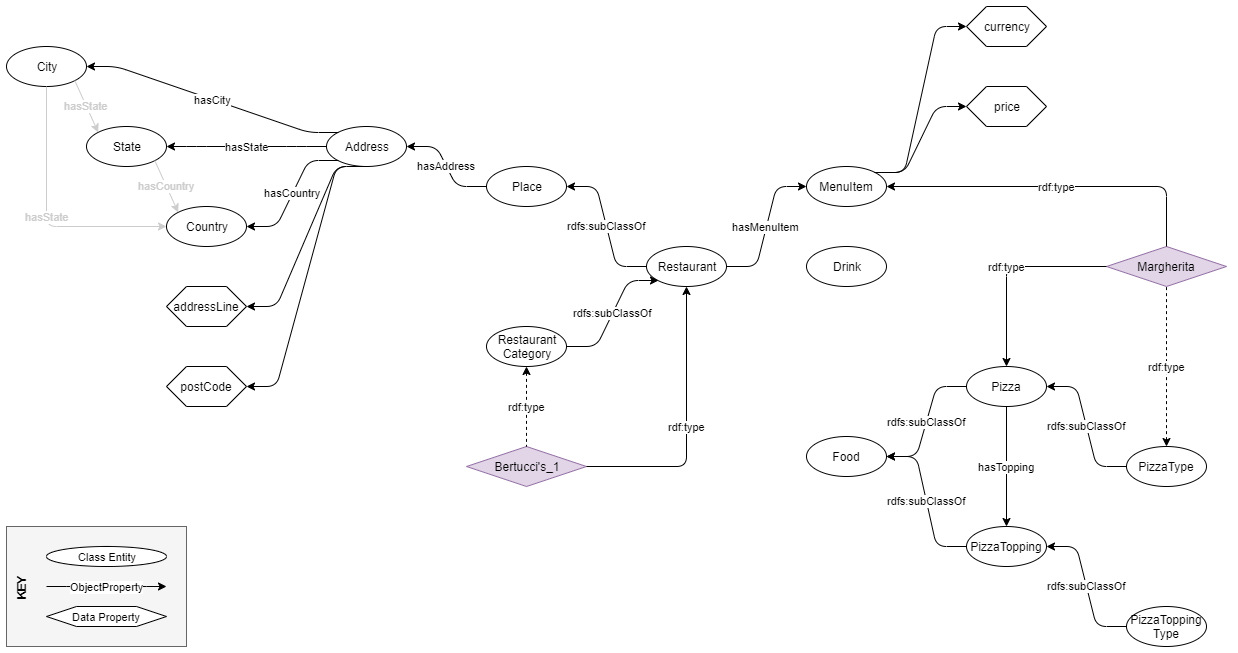


Figure 2. Ontology model for the tabular data

The complete list of entities created for the ontology, which will hereafter me referred to as *zdetor* is specified in Table 2.

|  |  |  |
| --- | --- | --- |
| **Classes** | **Object properties** | **Data Properties** |
| Place  Restaurant  RestaurantType  EthnicRestaurant  RestaurantType | hasAddress  hasMenuItem | name |
| Address  City  State  Country  Currency | hasCity  hasState  hasCountry  isAddressOf | postCode  addressLine |
| Drink  Food  Pizza  PizzaType  NamedPizza  PizzaType  PizzaTopping  PizzaToppingType | hasIngredient  hasTopping  isIngredientOf  isToppingOf |  |
| MenuItem | isMenuItemOf  hasCurrency | price  currency  description |

Table 2. Entities in the zdetor ontology

### Class Entities

#### Place

The Place class and subclasses are used to model the details in the name and category columns. Although we know that the places referenced in the spreadsheet are all restaurants (i.e. part of the *Restaurant* class) an overarching parent class of *Place* has been created in case in the future we want to add other types of places like, museums, galleries etc. that share common object properties (e.g. hadAddress). The Restaurants are split further into multiple subclasses based on the information found in the categories column. Some of the categories are linked to an ethnic menu therefore they’ve been grouped under the *EthnicRestaurant* subclass like (*MexicanRestaurant*, *ItalianRestaurant*, etc.). Other categories are based on food type (e.g. *SeafoodRestaurant*, *PizzaRestaurant*, etc.) therefore they’ve are direct subclasses of the *Restaurant* class. The subclasses are not disjoined as based on some category strings a restaurant can belong to multiple categories represented by subtypes.

#### Address / City / State / Country

Based on the data it was decided to make the *Address* an entity of each own rather than separate object properties of the *Place* class which would have been a simpler approach. However, having the *Address* being its own class, makes the model more flexible, in case an address house multiple restaurants (or more generally places). There are a few cases in the data we have that the same address houses two places:

e.g. ‘400 S Orlando Ave, 32751, FL’ is the address of ‘Nypd Pizza’ as well as ‘Francesco's Ristorante & Pizzeria’.

For city, state and country we only created classes in the zdetor ontology so that we can keep the original values from the spreadsheet as a name data property. Otherwise we wouldn’t need to create them in the ontology and we’d use the DBpedia URIs straight away. This is also the reason we’ve not created object properties to link cities to states and countries. In any case EDA has shown that the data is incorrect placing the same city in multiple different states.

#### Food

We’ve created a *Food* parent class, disjoined with a *Drink* class in order to group together anything that can be classified as food. For the specific dataset we assume that all the menu items are pizzas. We’ve broken down the *Food* class in the *Pizza* and *PizzaTopping* subclasses.

*Pizza* is further broken down to multiple *PizzaType* subclasses which are retrieved in a similar way like the *RestaurantTypes* from 2.2.1.1. Pizzas with a specific name are all grouped together under the parent class *NamedPizza* which is a subclass of *Pizza*. Like the restaurant categories, given the quality of the data, the subclasses of *Pizza* has not been set up as disjoined classes

*PizzaToppings* and the subclasses follow the same pattern as with the *Pizzas.*

#### MenuItem

The *MenuItem* has been set up as its own class so that we can have additional data properties like price, and currency. We decided to make this a separate class so that we can support in the future menu items that are not pizzas.

We do not need hasCity, hasState inside those entities

2.3.

I’ve created the classStringToURI = dict() as a dictionary of the format

{'class':{'string': uri}}

The first key will be used to denote the class we are creating a URI for and the second would be a dictionary of key value pairs. That is because we may have the same value for a restaurant and a pizza for instance so we do not want to mix the two.

Some of the Cities are not correct, Manchester… Universities… etc.

In order to decide create a unique URI for the address we concatenate the address like with the state column. That is because there is the case of the following two addresses that share the same address line but see to be different addresses altogether.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 222 E Main St | Mount Kisco | US | 10549 | NY |
| 222 E Main St | Collegeville | US | 19426 | Rahns |

For the categories we performed a basic NLP for keyword extraction in order to identify the most frequent terms in the ‘categories’ column. The insights from the NLP results informed the creation of some subclasses of the restaurant class. The list of subclasses is not exhaustive and it was created after manually inspecting the extracted keywords. For the NLP we allowed up to 3 word per expression and also allowed for the same word to appear in multiple expressions. We also allowed for the restaurant subclasses in the ontology to be NOT be disjoined since a lot of restaurants in the csv have multiple categories.

We follow a similar logic for the pizza subclasses

Reasoning

Originally made price a functional property

For the average price I grouped by currency as well. In this case where all prices are in USD that grouping is redundant. However if there were prices in other currencies as well then this grouping would ensure that only same currency prices would be summed up. In such a case we would get multiple results (i.e. one per currency).