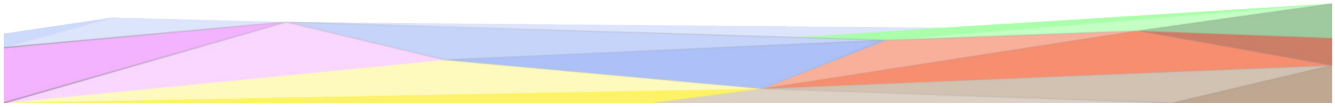
Dipartimento di Ingegneria e Scienza dell'Informazione

– KnowDive Group –

Go around JLU

Document Data:	Reference Persons:
2020.12.12	Chen Ziming,Du xiaolong,Gao xianjun,Wang yu(listed in no particular order)



Contents

Revision History:

Revision	Date	Author	Description of Changes

1 Knowledge Graph Codebook

The first of the two sections, in the current document, contains the codebook of the whole KG (Knowledge Graph), including the description of all the data and information that it contains.

1.1 Knowledge Graph general description

This sub section aims to give a general description of the KG, reporting:

- the context/domain in which the KG lives and works;
- *The Problem* the KG aims to solve;
- How the KG can solve *The Problem*

1.2 Data level

The data level section aims to describe in details the (final version of) datasets collected and managed by the KG, with a description of each variable involved.

1.2.1 Datasets general details

In this section are reported the metadata at datasets level, so the metadata regarding the sources, the authors, the collection methods, and so on.

1.2.2 Datasets metadata documentation

In this section are reported the metadata at dataset attribute level, through a description of each variable involved in the datasets collected, specifying the variable types, meanings, value-set (possible values), and every other meaningful variable information.

1.3 Ontology level

The ontology level section aims to describe the underlying KG ontology, through the description of its elements at each level, reporting so the language, conceptual and schema resources used within it.

1.3.1 Ontology general details

This first sub section of the ontology level description, report the general details such as authors, sources and the description of external ontology eventually adopted to generate the final one.

1.3.2 Ontology metadata documentation

In this section instead, are reported the more specific metadata describing the single elements of the ontology (terms, concepts, ETypes and relations).

1.4 Knowledge Graph Evaluation

In the final section of this first chapter, the KG Evaluation is reported. It aims to describe, through specific metrics, the quality of the overall KG on different aspect, like domain coverage, usability, domain representation, and other meaningful aspects.

2 Knowledge Graph Development Process

2.1 Scope Definition

2.1.1 Purpose

Nowadays, the lives of JLU students are too boring so that they need to find ways to entertain themselves. However, there are so many recreation places, saying, scenic spots, restaurants, cinemas that sometimes it is difficult to decide a route due to the complexity of querying data. Our goals is to build a DKG as an integration of knowledge and data required in the process of decision for JLU students, which means that by stating the theme, like birthday party or music concert, they can detect activities and matching scenic spots closely related with it. What's more they are capable of getting most suitable places for them based on the distance and price ranking.

2.1.2 Scope of work

Considering the amount of data we have found, size of the end application, time we are assigned, we have decided to integrate datasets, on the scope of only the city of Changchun, composed of points of interests, restaurants, shops, hotels, monuments, cinemas etc. to visit, their ratings and reviews from MeiTuan, and their distances from Amap.

2.1.3 Scenarios of usages

Name	Age	Usages	Description
Alen	20	Finding most suitable restaurants for him	Alen, as a college student, comes from Sichuan and he wants to eat Sichuan cuisines. However, as a college students, the budget and time is very limited. By using our system, Alen is able to find the restaurant that suit him most, which not only consider the price and time, but also recommend the best route for him to reduce the time cost for as much as possible.
Daniel	23	Using the system to find a route to entertain himself	Daniel is a freshman who has just arrived in Jilin University. This weekend he wants to go around Changchun and entertain himself. So by employing our system, Daniel is able to find a travel route which go through scenic spots. Prices, distances and specified time will be considered to recommend the best route for Daniel.

2.2 Inception

2.2.1 Competency Queries

In order to round up a collection of etypes and their properties,we will make use of the following table of competency queries as to what could possibly be question instances parameterized in terms of generic questions by persona and its systematized implementations.

Table 1: Definition of CQ

Persona	Num	Generalized Questions	Action
Alen	1.1	Find restaurants with Sichuan cuisines	Return a list of Sichuan restaurants depending on the ranking points.
Alen	1.2	Consider the factor of time and price to recommend Sichuan cuisines	Return a modified list of Sichuan restaurants at the limitation of certain consumer price range and time range.
Alen	1.3	Get the best route schedule	The system automatically generates a shortest route based on the chosen restaurants.
Daniel	2.1	Recommend senic spots based on prices and time	Return top recommended attractions based on rankings with the limitation of time and prices,
Daniel	2.2	Choose some spots	Return the detailed information(tags etc.) about those scenic spots
Daniel	2.3	Get the best route schedule	The system automatically generates a shortest route based on the chosen spots
Daniel	2.4	Adjust route schedule dynamically	return a final route given the chosen attractions and schedule .

Table 2: Additional information of CQ

NUM	TYPES	PROPERTIES
1.1-1.3	restaurant	type,location, consumer price range,business hour
2.1-2.2	tourist attractions	location,price,name,business hour
2.3-2.4	route schedule	A list of attractions, average cost, start time, end time, path.

2.2.2 Initial Datasets description

In this section, we decide to get data in the context of traveling around JiLin university. To obtain the location for entertainment, following websites are chosen as reference. Firstly, www.amap.com, by listing different destinations, the website will give us the destination locations, routes and corresponding prices. After completing the routes information, we should get the detailed information of the entertainment or restaurant from the website: cc.meituan.com. This website is a large comment website which can provide the detailed information about entertainment, restaurant, takeaway, cinema, etc. In this website, we can obtain the corresponding information which included price, brief introduction, comment, commented score, corresponding tag, etc. However, this website can't provide specific information when the users search for a key word. To deal with the problem, we will search for information in the website:www.ctrip.com. In this website, we can search corresponding information with a specify tag.

2.2.3 Datasets collection process

In this section, we will introduce the datasets procession. Though it's not difficult to collect the target information, the whole procession still needs large amount of work. In order to decrease the workload, we write a script in Python to obtain the information. This script includes the information obtaining function which can be used to obtain the corresponding information. Take getting tourist reviews as an example, in this section, we first get the designated website which is can be changed by the destination id. After that, we will get the source code of the website and will locate the section which records the tourist information by the regular expression. Because the information is constructed by json language, by searching the specified tag, we can obtain the information including the comment and score. By this way, all the function can be constructed and used to get information.

2.2.4 Inception level evaluation

The last section of the Inception phase report the evaluation of the outcomes obtained in this phase, through specif evaluation metrics.

2.3 Informal Modeling

2.3.1 Schema level

The schema level in this phase report the first informal definition of the ETypes and of the EER model constructed using them.

2.3.1.1 ETypes and EER Model definition

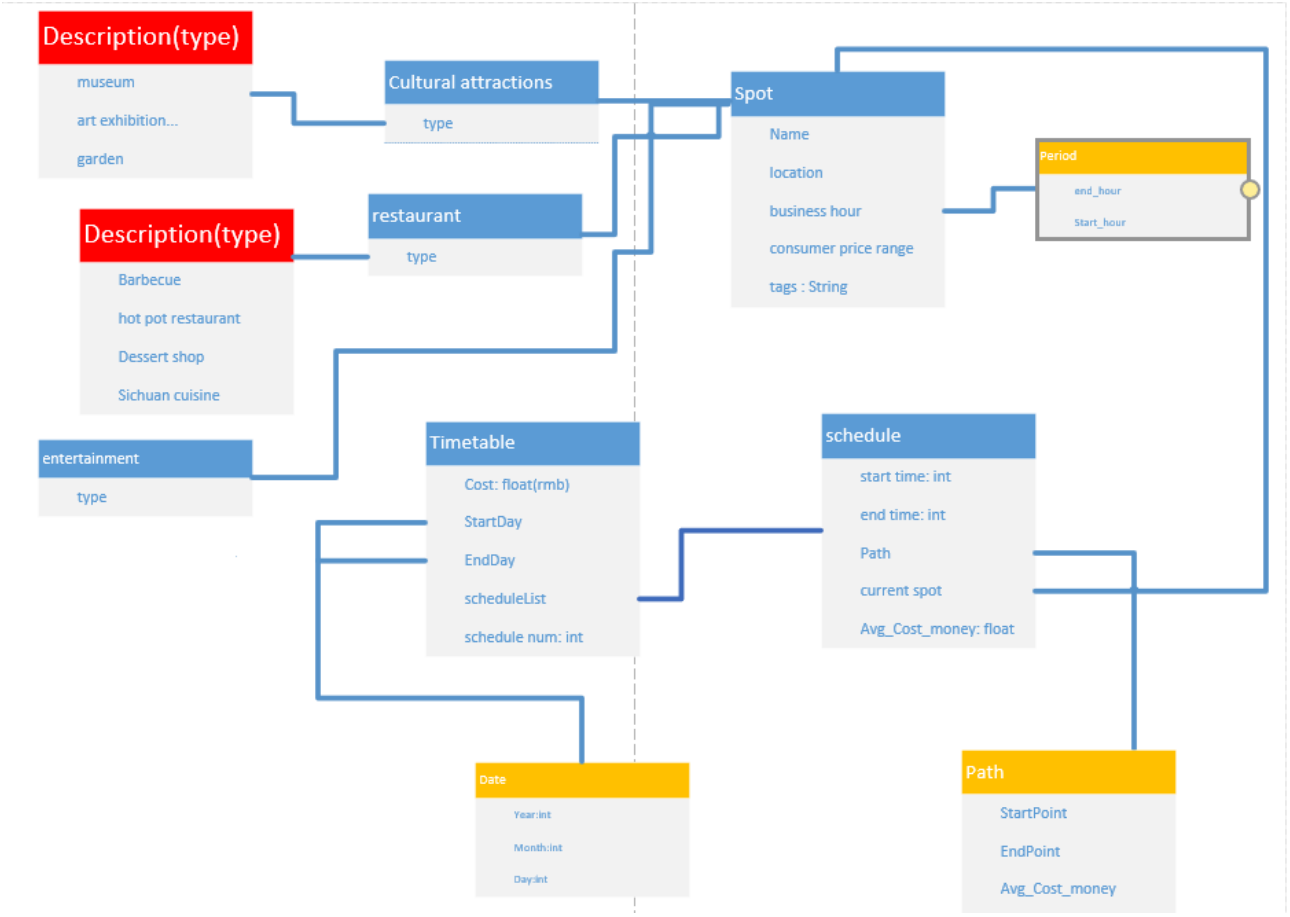


Figure 1: EER Diagram of the data used inside the project

2.3.1.2 Variance respect CQs definition

Our Core entities are timetable, schedule and spots.

Timetable: It represents a recommended travel schedule generated by our software automatically .It has five attributes.

- cost: the predicated total cost of the tour (dt: float);
- StartDay: the start date of the tour (dt: date);
- EndDay: the end date of the tour (dt: date);
- scheduleList(*): a list of schedule (dt: list);
- schedule num(*): the length of the scheduleList (dt: int);

Schedule: It represents an activities at a particular place and a particular time. It has five attributes.

- start time: the start time of the activities (dt: date);

- end time: the endtime of the activities (dt: date);
- Path(*): Defines how to reach the designated destination ;
- Avg_Cost: the average of cost in this spot (dt: float);

Spot: It represents a Service place . It has three children and six attributes.

- Name: the name of the spot (dt: string);
- location: the location of the spot (dt: string);
- Business hour(*): Defines the business hours of service place (dt: period) ;
- consumer price range: Average consumption of customers in service places (dt:float);
- Style and atmosphere: Specific types of service places, such as hotpot restaurant,
- barbecue shop, etc (dt: enum);
- small_spot: Son spot,In some cases it can be null;

2.3.2 Data level

The data level section in this phase reports the evolution of the datasets collected previously, reporting the metadata information for each new data, or new version of data, obtained.

2.3.2.1 Datasets management process

During the Informal Modeling phase the datasets collected in the previous phase are managed in order to obtain more suitable sets of data. As we know the datasets collected from various resources can be really dirty. It will have both relevant and irrelevant information. In short, it is wise to get rid of the useless information that we have in our datasets. The datasets we have are in Json formats which can be easily handled by using Python programming language. We used related package to manipulate the structured datasets in a dataframe and delete the irrelevant attributes. The resultant dataframe was again stored as a CSV in the repository so that it can be used as an input in the upcoming phase.

2.3.2.2 Datasets metadata documentation

Table 3: MeiTuanOpenData.json

Information	Description
Dataset Owner	Beijing Sankuai Online Technology Co., Ltd.
Dataset Publisher	Meituan Technology Co., Ltd.
Publish Date	2010.3.4
Dataset Language	English&&Chinese
Coverage Range	China
Type of Website	Group buying website
URL	http://www.meituan.com/
Format	Json

Table 4: AmapData.json

Information	Description
Dataset Owner	Gaode Software Co., Ltd.
Dataset Publisher	Gaode Software Co., Ltd.
Publish Date	2011.5.6
Dataset Language	English&&Chinese
Coverage Range	China
Type of Website	Travel
URL	http://ditu.amap.com/
Format	Json

2.3.2.3 Variance respect Inception datasets

In this section, we will use data from different website. From the route information, we need to collect information including different routes to satisfy different request, prices based on the routes, the locatiosn of the destination and the distances. For the destination information, we need to collect the type which can define the basic type, the average price for each person, the comment, and the menu if it is a restaurant, the tag which can describe its feature, the time when it's open, the score by others, etc. These variances will be used to construct the knowledge graph.

2.4 Formal Modeling

This section is dedicated to the Formal Modeling phase description. The Section is divided in Schema and Data level in order to report the details regarding both the ontology generated and the datasets version in the current phase.

2.4.1 Schema level

The schema level section in the current phase, reports the detailed description of the ontology generation.

2.4.1.1 Ontology definition

General The first step of the process for the development of the ontology schema consist on searching for other reference ontologies. We found one valid ontology from the DBpedia.Then we compare the names on our ER with the ones already present on UKC to make the our ER more explicit and standardized. Based on this we developed a list of entities to add to protégé and exported to owl / XML file.

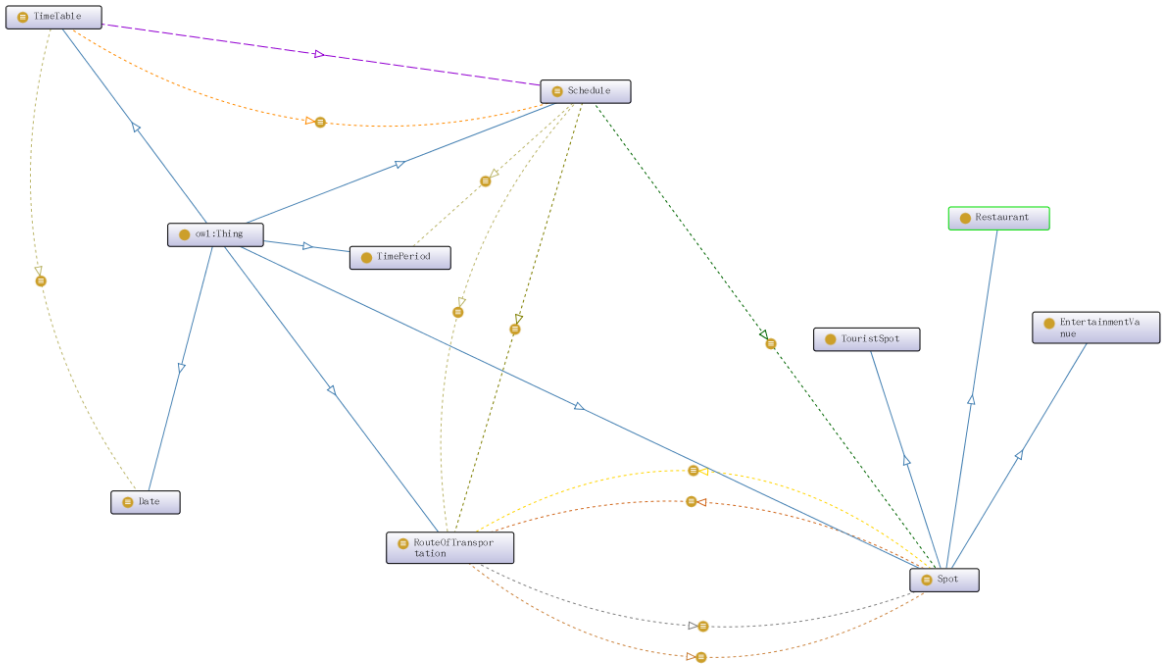


Figure 2: Pictures generated by protege

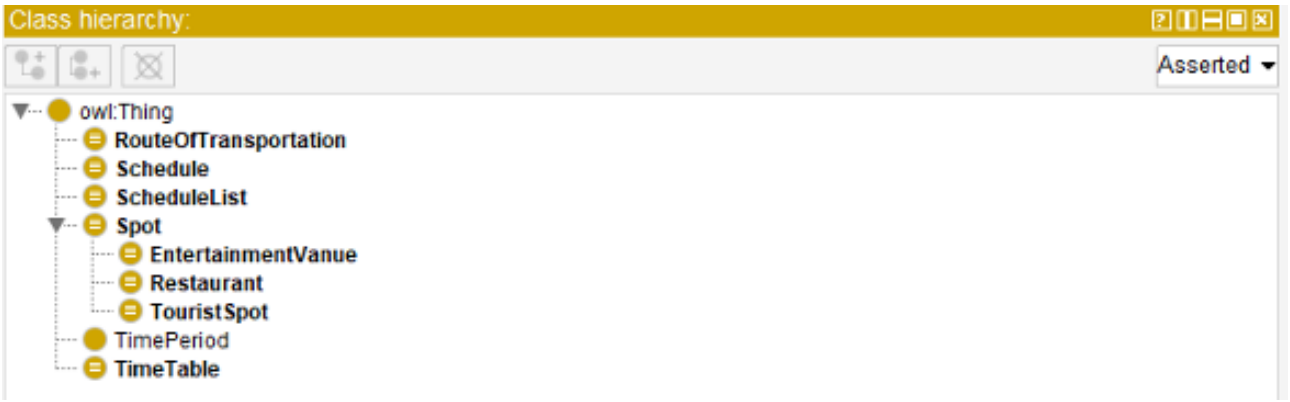


Figure 3: Classes

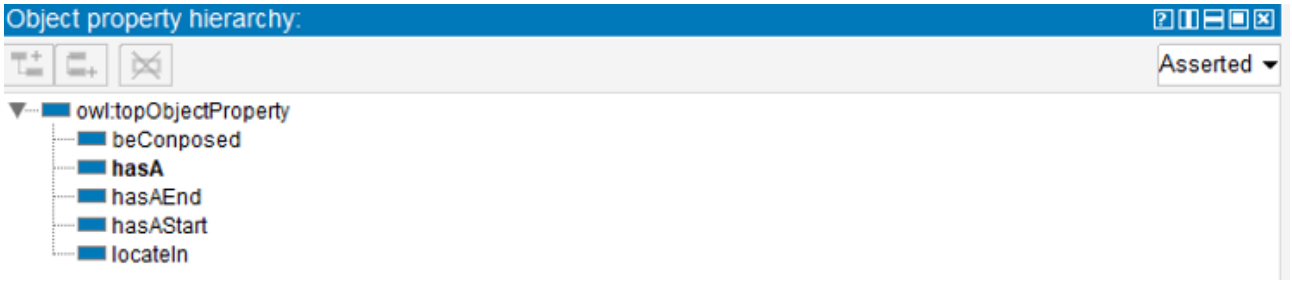


Figure 4: Data Property

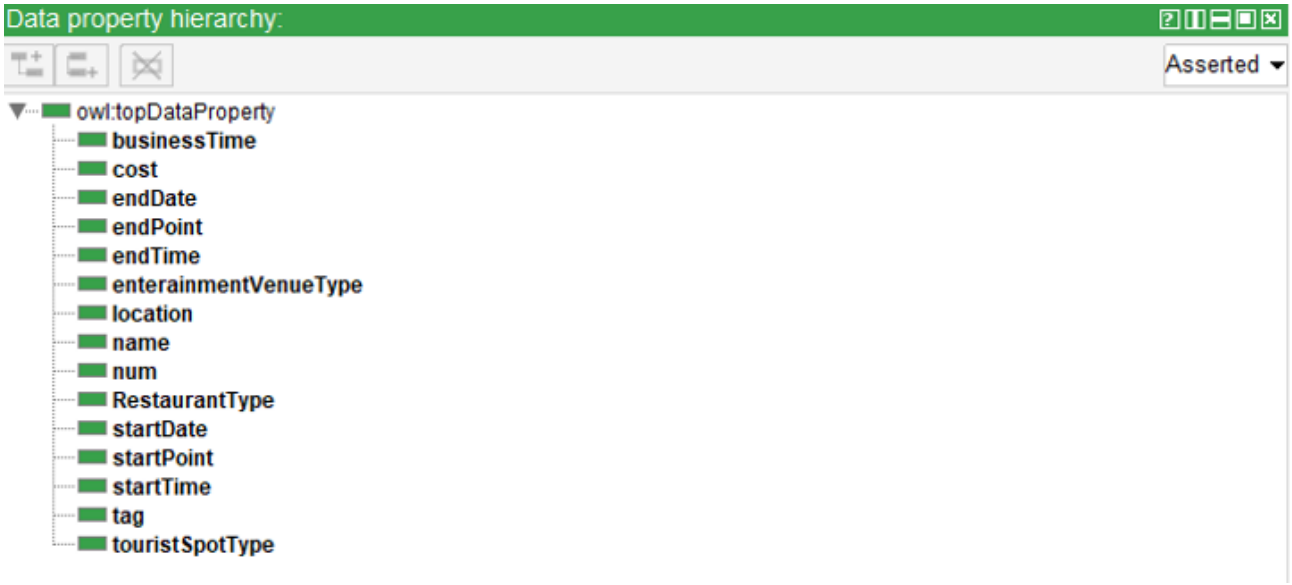


Figure 5: Object Property

2.4.1.2 Variance respect to the EER Model

Core entities: Our Core entities are timetable, schedule and spot.

TimeTable: it represents a recommended travel schedule generated by our software automatically .It has three attributes.

- Cost: the predicated total cost of the tour (dt: float);
- StartDateTime: the start time of the tour (dt: string);
- EndDateTime: the endtime of the tour (dt: string);

Schedule: it represents an activities at a particular place and a particular time. It has four attributes.

- StartDateTime: the start time of the activity (dt: string);

- EndDateTime: the endtime of the activity (dt: string);
- Name: name of current spot(dt: string);
- Average: the average of cost in this spot (dt: float)

Spot: it represents a Service place . It has three children(Restaurant, TouristSpot and EntertainmentVenue) and four attributes.

- Name: the name of the spot (dt: string);
- Location: the location of the spot (dt: string);
- Average: Average consumption of customers in service places
- (dt:float);
- Tags: Specific types of service places, such as hotpot restaurant, barbecue shop, etc .(dt: enum);

Common entities: Our common entities are taken from Dbpedia, and are the RouteOfTransportation.

- RouteOfTransportation: it represents It defines the starting and ending place and distance of a route. It has three attributes.;
- StartPoint: the start of the route(dt: string); EndPoint: the end of the route(dt: string);
- Distance: Distance between start and end(dt:float);

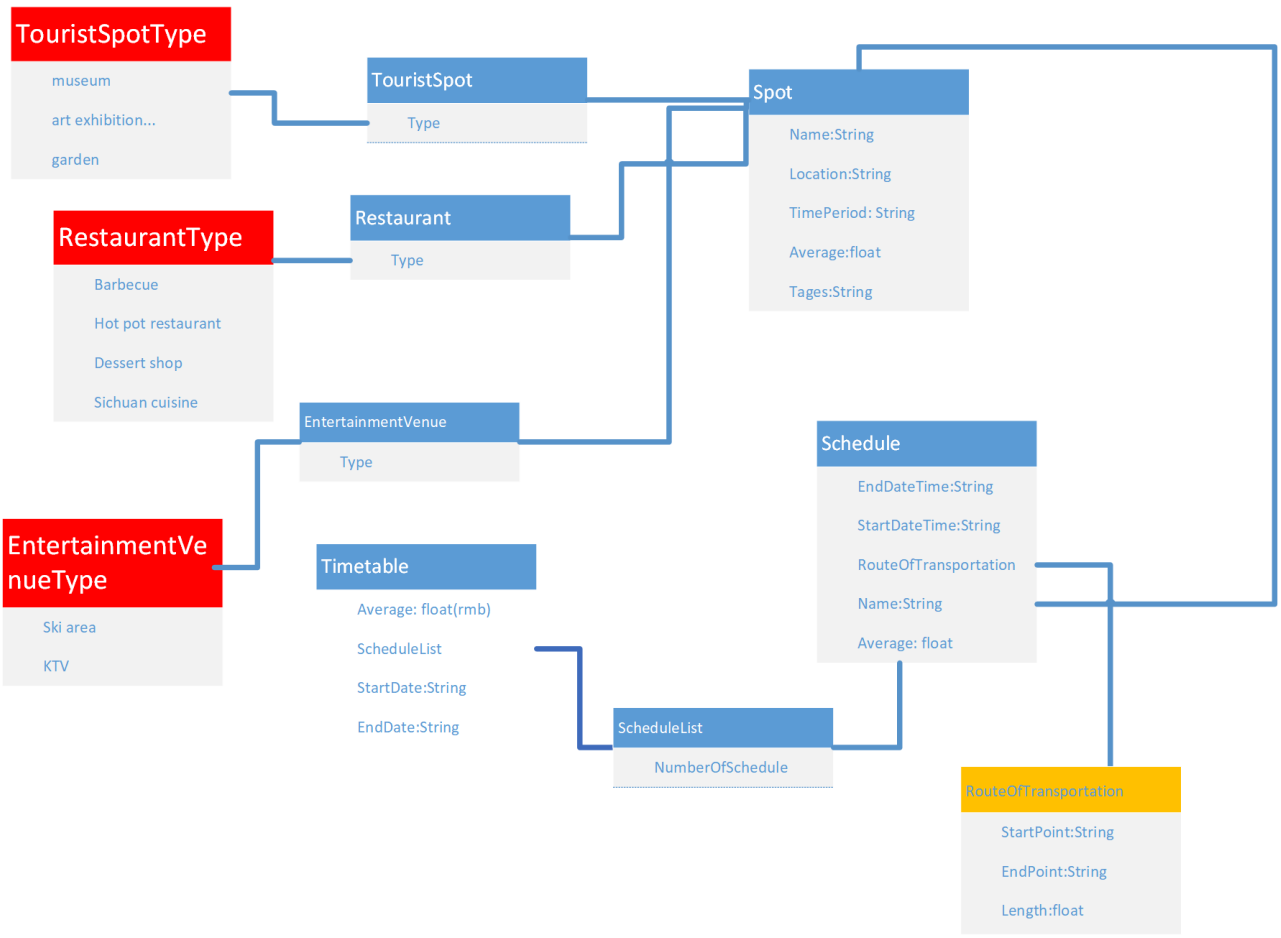


Figure 6: EER Diagram of the data used inside the project

2.4.2 Data level

As in the previous phase the data level section here, reports the description of the new version of the datasets, after formatting operations.

2.4.2.1 Formal Modeling datasets management

During the Informal Modeling phase the datasets collected in the previous phase are managed in order to obtain more suitable sets of data. As we know the datasets collected from various resources can be really dirty. It will have both relevant and irrelevant information. In short, it is wise to get rid of the useless information that we have in our datasets. The datasets we have are in Json formats which can be easily handled by using Python programming language. We used related package to manipulate the structured datasets in a dataframe and delete the irrelevant attributes. The resultant dataframe was again stored as a CSV in the repository so that it can be used as an input in the upcoming phase.

2.4.2.2 Datasets metadata documentation

Table 5: MeiTuanOpenData.json

Information	Description
Dataset Owner	Beijing Sankuai Online Technology Co., Ltd.
Dataset Publisher	Meituan Technology Co., Ltd.
Publish Date	2010.3.4
Dataset Language	English&&Chinese
Coverage Range	China
Type of Website	Group buying website
URL	http://www.meituan.com/
Format	Json

Table 6: AmapData.json

Information	Description
Dataset Owner	Gaode Software Co., Ltd.
Dataset Publisher	Gaode Software Co., Ltd.
Publish Date	2011.5.6
Dataset Language	English&&Chinese
Coverage Range	China
Type of Website	Travel
URL	http://ditu.amap.com/
Format	Json

2.4.2.3 Variance respect Informal Modeling datasets

In this section, we will use data from different website. From the route information, we need to collect information including different routes to satisfy different request, prices based on the routes, the locatiosn of the destination and the distances. For the destination information, we need to collect the type which can define the basic type, the average price for each person, the comment, and the menu if it is a restaurant, the tag which can describe its feature, the time when it's open, the score by others, etc. These variances will be used to construct the knowledge graph.

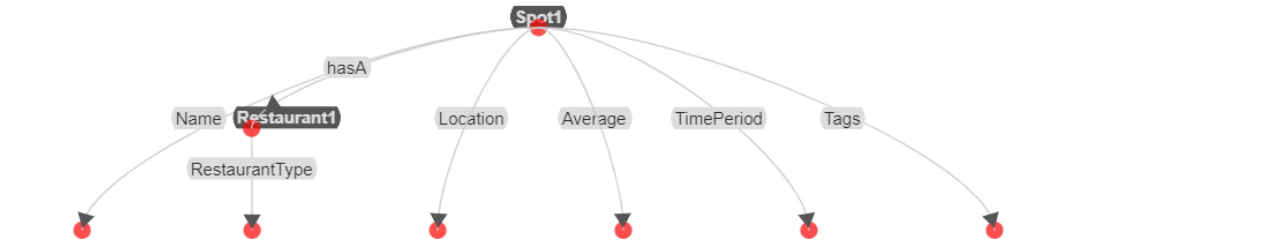
2.5 Data integration

This section is dedicated to the Data Integration phase description.

2.5.1 Data integration operations and tool

This section is dedicated to the description of the usage of the data integration tool that allows to map the datasets generated and well formatted in the previous phases, with the final ontology generated. The last datasets adaptation performed using the tool, as well as the mapping operation are here detailed.
During the final data-integration,we used the tool named karma to integrate the owl file and the csv file.Firstly,two files were imported into karma.Secondly, map each data label in the csv file to its corresponding property and class

in the owl file. After that, set its foreign key for each class in the owl file. Then complete the mapping between the data in the csv file and the ontology in semantic in owl



name	type	location	price	time	tags
湘肴世家湘菜馆	川菜	净月开发区永顺路与樱花东街交叉口伟峰东域下沉式商业广场A8-118号商铺	76	10:00-21:00/Mon-Sun	null
川锦汇自选麻辣拌（巴黎春天百货店）	川菜	朝阳区红旗街红旗街西一胡同南行100米	10	9:00-21:00/Mon-Sun	null
弘扬川味	川菜	汽车开发区锦程大街7号天吉锦城苑5幢2单元103号房	34	08:30-20:30/Mon-Sun	null
吉祥富贵重庆烧鸡公水煮鱼	川菜	朝阳区西朝阳路与云鹤胡同交汇	38	10:00-22:00/Mon-Sun	null
渝府干锅辣鸭头	川菜	双阳区柏山路东鹿城中央区第3幢115号门市（金柜音乐广场对面）	38	08:30-21:30/Mon-Sun	null
蜀道川厨	川菜	南关区北海路与珠海南胡同交汇康诺口腔北行100米（高丽串城北侧 北海路口）	45	09:30-21:30/Mon-Sun	null
铭苑麻辣香锅-重庆烤鱼	川菜	宽城区青年路住邦城市广场四楼	44	10:00-22:00/Mon-Sun	null
铭苑麻辣香锅	川菜	朝阳区光明路与新华	25	09:30-21:00/Mon-	null

Figure 7: The final result of data integration

2.5.2 Variance respect Formal Modeling datasets

The last section of the data integration phase aims to describe the variance, analyzing the differences, between the datasets integrated with the ontology, in the data integration platform which contain the KG, and the datasets collected in the previous phase. This analysis can highlight the results of the operations performed during the final phase of the data integration process.