



UNIVERSITY  
OF TRENTO - Italy



DIPARTIMENTO DI INGEGNERIA E SCIENZA DELL'INFORMAZIONE

– KNOWDIVE GROUP –

# Go around JLU

---

Document Data:

2020.12.12

(listed in no particular order)

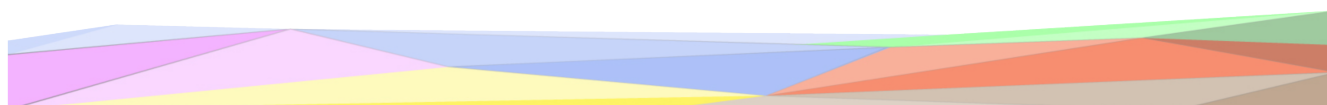
Reference Persons:

Chen Ziming, Du xiaolong, Gao xianjun, Wang yu

© 2020 University of Trento

Trento, Italy

KnowDive (internal) reports are for internal only use within the KnowDive Group. They describe preliminary or instrumental work which should not be disclosed outside the group. KnowDive reports cannot be mentioned or cited by documents which are not KnowDive reports. KnowDive reports are the result of the collaborative work of members of the KnowDive group. The people whose names are in this page cannot be taken to be the authors of this report, but only the people who can better provide detailed information about its contents. Official, citable material produced by the KnowDive group may take any of the official Academic forms, for instance: Master and PhD theses, DISI technical reports, papers in conferences and journals, or books.



---

# Contents

<b>1</b>	<b>Knowledge Graph Codebook</b>	<b>1</b>
1.1	Knowledge Graph general description . . . . .	1
1.2	Data level . . . . .	1
1.2.1	Datasets general details . . . . .	1
1.2.2	Datasets metadata documentation . . . . .	1
1.3	Ontology level . . . . .	1
1.3.1	Ontology general details . . . . .	1
1.3.2	Ontology metadata documentation . . . . .	1
1.4	Knowledge Graph Evaluation . . . . .	2
<b>2</b>	<b>Knowledge Graph Development Process</b>	<b>2</b>
2.1	Scope Definition . . . . .	2
2.1.1	Purpose . . . . .	2
2.1.2	Scope of work . . . . .	2
2.1.3	Scenarios of usages . . . . .	2
2.2	Inception . . . . .	3
2.2.1	Competency Queries . . . . .	3
2.2.2	Initial Datasets description . . . . .	4
2.2.3	Datasets collection process . . . . .	4
2.2.4	Inception level evaluation . . . . .	4
2.3	Informal Modeling . . . . .	4
2.3.1	Schema level . . . . .	4
2.3.2	Data level . . . . .	6
2.3.3	Informal Modeling Evaluation . . . . .	7
2.4	Formal Modeling . . . . .	7
2.4.1	Schema level . . . . .	7
2.4.2	Data level . . . . .	8
2.4.3	Formal Modeling Evaluation . . . . .	8
2.5	Data integration . . . . .	8
2.5.1	Data integration operations and tool . . . . .	8
2.5.2	Variance respect Formal Modeling datasets . . . . .	9

## 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 KDG 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. Apart from, the function of automatically generation of Facebook or Tweets is also inserted into our system.

### 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](http://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](http://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](http://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





- 
- end time: the endtime of the activities (dt: date);
  - Path(\*): Defines how to reach the designated destination ;
  - spot(\*):
  - 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	<a href="http://www.meituan.com/">http://www.meituan.com/</a>
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	<a href="http://ditu.amap.com/">http://ditu.amap.com/</a>
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 location 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.

### 2.3.3 Informal Modeling Evaluation

## 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**

This section reports in details how the ontology is generated starting from the informal schema of the previous phase, which tools are used to do that, as well as usage of external ontology resources adopted to obtain the final KG ontology. Moreover a list of metadata is reported in this section, in order to describe all the elements of the ontology defined.

#### **2.4.1.2 Variance respect to the EER Model**

Once the ontology has been built, this section report the differences, and so the variance, respect the EER model defined in the previous phase. This a way to define the quality of the outcomes for the current phase as well as the alignment of the overall project development process.

### **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**

In this section are reported the operations and the tools adopted to format the dataset collected, in order to align them to the ontology definitions generated at schema level.

#### **2.4.2.2 Datasets metadata documentation**

In this section eventually new metadata information are added in order to describe the evolution of the datasets.

#### **2.4.2.3 Variance respect Informal Modeling datasets**

This section aims to define the variance between the data elements (datasets and attributes within them) produced in this phase, and the initial datasets collected in the previous phase. This a way to define the quality of the outcomes for the current phase as well as the alignment of the overall project development process.

### **2.4.3 Formal Modeling Evaluation**

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

## **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.

---

### 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.