Dipartimento di Ingegneria e Scienza dell'Informazione

– KnowDive Group –

Codebook Template

Document Data:	Reference Persons:
- date -	- authors -

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

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

The second chapter of this document aims to describe, in a detailed way, the KG development process. The sections below describe each phase of the KG building project, reporting for each phase, the description of the datasets and their evolution respect the previous phases, the schema construction which will generate the KG ontology in the end, as well as the description of the procedures adopted to manage the data and finally achieve those results. Moreover for each phase is reported an evaluation section, which aims to evaluate the quality of the results achieved at the end of each phase.

2.1 Scope Definition

This section aims to define the purposes for the creation of the Knowledge Graph, describing the context in which it has to live as well as the usage scenarios in which it can be involved. Moreover a list of general questions regarding the objectives to achieve through the development of the Knowledge Graph, is reported here. More in general this section has to give all those information which allow to understand which is the problem to solve, and why we need a KG to solve it.

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
Allen	20	Detecting gym and restaurants that can help him keep fit	Allen is shy and he prefers to stay alone, so he needs to find gym which is quiet and has only a few people in it. So he can use our system to detect unpopular gym and restaurants selling fitness meal.
Zachary	26	Using the system to identify hotspot suitable for chatting with friends	Zachary, as stated in the interest, likes to chat with friends while eating hot pot, but many hotpot shops will limit the time, and the atmosphere is also very important. So through the system he can find time-unlimited hotspot to enjoy with friends.
Daniel	25	Searching places that can contain more than 10 people for singing and dancing	Daniel has just finished all his exams and want to release himself with classmates, through our system, they can find places with enough room to have fun.
Mike	23	Finding a places to date with girlfriend	By employing the system they can detect romantic and peaceful places to date, like scenic spots or parks.

2.2 Inception

This section is dedicated to the Inception phase description. Here are reported the initial definitions for CQs (Competency Queries), initial datasets collected and the relative metadata.

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.

Persona	Num	Generalized Questions	Action
Allen	1.1	Give the list of fitness place with business hours that match his schedule	Return a list of fitness place ordered by distance given the current date and place
Allen	1.2	Process the list of fitness place with private and undisturbed environment	Return the list of fitness place with specific style and atmosphere
Allen	1.3	Process the list with his favorite fitness equipment	Return a list of fitness place with specific equipment category
Allen	1.4	Process the list with a certain level of consumption	Return a list of fitness place given a certain consumer price range
Allen	1.5	Give a list of restaurants nearby JLU	Return a list of restaurants ordered by distance Given origin
Allen	1.6	Process a list of sichuan restaurants	Return a list of restaurants given specific cuisine
Allen	1.7	Process a list of restaurants which support takeaway and a certain level of consumption	Return a list of restaurants given a certain consumer price range and specific features(takeaway)
Daniel	2.1	Give a list of recommended attractions nearby JLU during the New Year's Day	Return a list of recommended attractions informationn (name, distance, price, characters) given specific date
Daniel	2.2	Choose a favorite spot	Return a specific scenic spot and show the detailed scenery in the scenic spot (picture, small scenic spot)
Daniel	2.3	Give a list of recommended route schedule(including budget)	The system automatically generates a better route schedule given a specific location
Daniel	2.4	adjust final a route schedule	return a final route given the chosen attractions and schedule .

NUM	TYPES	PROPERTIES
1.1-1.4	Fitness place	name,location,business hour,style and atmosphere,equipment category,consumer price range
1.5-1.7	restaurant	type,specific cuisine,features (location, consumer price range,support take away or not)
2.1-2.2	tourist attractions	location,characters,price,name ,business time, specific scenic spot
2.3-2.4	route schedule	pairs of the time period and the corresponding attractions, total cost

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

This section is dedicated to the Informal Modeling phase description. The Section is divided in Schema and Data level in order to report the details of the elements involved in the generation of the schema, as well as the description of the datasets evolution in this phase. Moreover a specif section, one for each level, reports the difference between the elements defined in this phase and the definitions in the previous phase, analyzing in this way the variance in the different phases.

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

This section reports an informal definition of the ETypes involved in the datasets collected in the previous phase. This section includes a list of metadata associated to each of the elements generated.

2.3.1.2 Variance respect CQs definition

This section aims to define the variance between the schema elements produced in this phase, and the definition of the CQs reported 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.

2.3.2 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 filtered and managed in order to obtain more suitable sets of data. In this section are described the procedures adopted to obtain that result.

2.3.2.2 Datasets metadata documentation

In this section is reported a list of new metadata in order to describe the modification performed on each datasets and attribute, to achieve the new version of the datasets.

2.3.2.3 Variance respect Inception 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.

2.3.3 Informal Modeling Evaluation

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

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 stating 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.