UNIVERSITY
OF TRENTO - Italy

Dipartimento di Ingegneria e Scienza dell'Informazione

– KnowDive Group –

KGE 2023 - Project Report Template

Document Data:	Reference Persons:
September 14, 2023	Author1, ..., AuthorN

Index:

Revision History:

Revision	Date	Author	Description of Changes
0.1	September 14, 2023	Author1	Document created

1 Introduction

Reusability is one of the main principles in the Knowledge Graph Engineering (KGE) process defined by iTelos. The KGE project documentation plays an important to enhance the reusabiltiy of the resources handled and produced during the process. A clear description of the resources as well as of the process (and sub processes) developed, provides a clear understanding of the project, thus serving such an information to external readers for the future exploitations of the project's outcomes.

The current document aims to provide a detailed report of the project developed following the iTelos methodology. The report is structured, to describe:

- Section 2: Definition of the project's purpose and its domain of interest.
- Section 3: High level description of the project development, based on the two main sub process considered by iTelos, producer and consumer, respectively.
- Sections 4, 5, 6, 7 and 8: The description of the iTelos process phases and their activities, divided by knowledge and data layer activities, as well as considered from the point of view of the producer first, and the consumer later.
- Section 9: The description of the evaluation criteria and metrics applied to the project final outcome.
- Section 10: The description of the metadata produced for all (and all kind of) the resources handled and generated by the iTelos process, while executing the project.
- Section 11: Conclusions and open issues summary.

2 Purpose and Domain of Interest (DoI)

Purpose:

To create a reusable knowledge graph that accurately represents the network of bus routes in Ulaanbaatar city. This involves connecting various bus stops throughout the city to establish a visualization of the public transportation system. By connecting these routes and stops, the knowledge graph will serve as a valuable resource for urban planners, public transportation planners for decision making, optimizing routes, and enhancing overall efficiency and accessibility withing Ulaanbaatar's transportation system.

3 Project Development

This section describes, at top level, how the project's purpose will be satisfied. More in details the current section is divided in two main subsections, defined as follows.

3.1 Data Production

The description of which (quality) data needs to be created to satisfy the project purpose. In this sub-section the role of the data producer is central. The sub-section aims at describing how the data producer enables the subsequent work of the data consumer, by creating the data required to satisfy the project's purpose.

3.2 Data Composition

This sub-section aims at describing the work of the data consumer in the project. More in details, how the consumer composes the data, previously created by the producer, with the objective of creating a Knowledge Graph suitable to satisfy the project's purpose.

4 Purpose Formalization

4.1 Scenarios definition

In life, people often encounter situations where they need to choose a bus route that goes to their desired destination. When this happens, finding the most useful route becomes particularly significant.

Scenario 1:
Planning Bus Routes: Someone who plans bus routes uses the knowledge graph to decide where buses should go, based on where people live and where they need to go.

Scenario 2:
From Sansar to the NUM by bus.

Scenario 3:
From Nalaikh to the city center by bus.

4.2 Personas

The characters involved are these:
Public transport companies: They run the bus and they plan the routes including paths.
Public transport passengers: They take the public bus from starting point to their destination.

Persona 1:
Bayaraa is a transportation planner in Ulaanbaatar.

Persona 2:
Tsetseg is a 21-year-old student living in Sansar. She goes to NUM mostly every day.

Persona 3:
Bold is a 62-year-old living in Nalaikh district. He sometimes goes to Ulaanbaatar city center to meet his children

4.3 Competency Questions (CQs)

CQ1:
Bayaraa optimizes a bus route for efficiency and effectiveness.

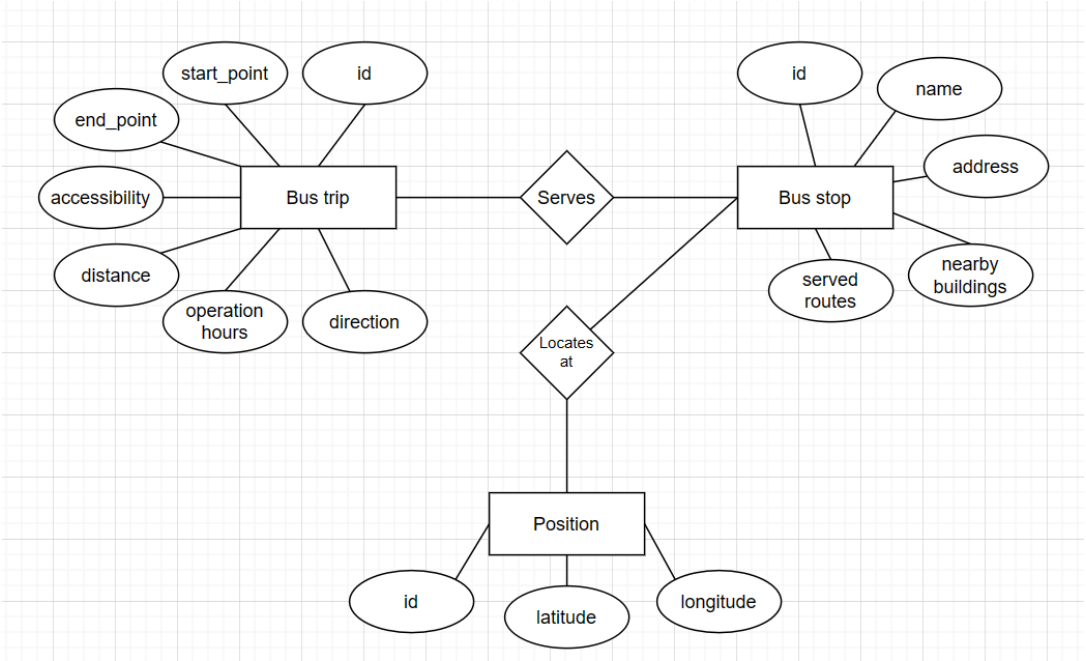
CQ2:
Tsetseg was at her grandparents' house on the weekend. She goes to the university from Yarmag.

CQ3:
Bold comes to visit his son who lives in Zaisan.

4.4 Concepts identification

Scenarios	Personas	Competency Questions	Entities	Properties	Focus	Popularity
1	1	1	Bus trip	id, start_point, end_point, distance, operating_hours, direction,accessibility,	Core	Core
2,3	2,3	2,3	Bus stop	id, name, address, nearby_buildings, served_routes	Common	Common
2,3	2,3	2,3	Position	id, latitude, longitude	Common	Common

4.5 ER model definition



5 Information Gathering

This section aims at reporting the execution of the activities involved in the Information Gathering iTelos phase. The report, starting from the current section, is organized along two main dimensions. The first one considers the parallel execution of the producer and consumer processes, while the second dimension takes into account the activities operating over data and knowledge layers.
In this section are described both the resourced selected, and the sources from which such resources have been retrieved.

Information Gathering sub activities:

- **Producer activities**: these activities aim at collecting ”informal” resources from sources with an higher level of heterogeneity. The resources collected by the producer process are not compliant with the iTelos quality and reusability guidelines. Those are the resources that the producer will transform into quality resources at the end of the process.
  - Knowledge layer:
    * Sources description
    * Informal resources collection and scraping;
    * Informal resources classification over common, core and contextual
  - Data layer:
    * Sources description
    * Resources collection and scraping;
    * Resources classification over common, core and contextual
- **Consumer activities**: these activities aim at collecting the already available resources considered for the project. More in detail the resources here described, are ”quality and formal” resources (compliant with the quality and reusability guidelines defined by iTleos. 6*, or at least 5*) which don’t need to be processed or created by a data producer. The resources described in this section are those that can be already composed by the data consumer to satisfy the project’s purpose.
  - Knowledge layer:
    * Sources description

    - Formal resources collection;
    - Formal resources classification over common, core and contextual
  - Data layer:
    - Sources description
    - Formal resources collection;
    - Formal resources classification over common, core and contextual

The report of the work done during the first phase of the methodology, has to includes also the description of the different choices made, with their strong and weak points. In other words the report should provide to the reader, a clear description of the reasoning conducted by all the different team members.

6 Language Definition

This section is dedicated to the description of the Language Definition phase. Like in the previous section, it aims to describe the different sub activities performed by all the team members, as well as the phase outcomes produced. Moreover, like the previous section, the organization of this section follows the two dimensions based on producer and consumer roles, and the two layers of activities, knowledge and data, respectively.

Language Definition sub activities:

- **Producer activities**: these activities aim at fixing the language (concepts and words) used to represent the information required to satisfy the project purpose. With this objective, the knowledge ad data resources are handled in this phase following the below activities:
  - Knowledge layer:
    - Concept identification
    - UKC alignment
  - Data layer:
    - Dataset filtering
- **Consumer activities**: the resources considered by the consumer are already formalized bya producer process. Nevertheless, new concepts, words and meanings can be consideredby the consumer process by considering the composition instead of the production of single resources. For this reason the below activities have to be described form the consumer point of view.
  - Knowledge layer:
    - Concept identification
    - UKC alignment
  - Data layer:

        * Dataset filtering

The report of the work done during this phase of the methodology, has to includes also the description of the different choices made, with their strong and weak points. In other words the report should provide to the reader, a clear description of the reasoning conducted by all the different team members.

7 Knowledge Definition

This section is dedicated to the description of the Knowledge Definition phase. Like in the previous section, it aims to describe the different sub activities performed by all the team members, as well as the phase outcomes produced. Moreover, like the previous section, the organization of this section follows the two dimensions based on producer and consumer roles, and the two layers of activities, knowledge and data, respectively.

Knowledge Definition sub activities:

- **Producer activities**: these activities aim at defining the knowledge structure of the information to be considered to satisfy the project purpose. More in details, the producer process, in this phase, aims at defining the knowledge structure for each dataset to be formalize, singularly. The data within such datasets, are then aligned with the structure d knowledge.
  - Knowledge layer:
    * Teleontology definition
    * Teleology definition
  - Data layer:
    * Dataset cleaning and formatting
- **Consumer activities**: the consumer activities have the same objectives like those executed by the producer before. Nevertheless, the consumer process define the knowledge structure, and perform the data alignment, over the composition of more datasets instead of the treating them singularly.
  - Knowledge layer:
    * Teleontology definition
    * Teleology definition
  - Data layer:
    * Dataset cleaning and formatting

The report of the work done during this phase of the methodology, has to includes also the description of the different choices made, with their strong and weak points. In other words the report should provide to the reader, a clear description of the reasoning conducted by all the different team members.

8 Data Definition

This section is dedicated to the description of the Data Definition phase. Like in the previous section, it aims to describe the different sub activities performed by all the team members, as well as the phase outcomes produced. Unlike the previous section, the organization of the current one follows a single dimension, the one considering the distinction between producer and consumer processes. The division between knowledge and data activities in this section is not defined, because, in this phase the two layers are merged to form a single data structure composed by the knowledge structures defined in the last section, and the aligned dataset. The obtained result is a structured Knowledge Graph including both the two layers.
Data Definition sub activities:

- **Producer activities**: the producer activities aim at merging the knowledge layer of a single dataset with the data values present within such a dataset.
  - Entity identification
  - Data mapping
- **Consumer activities**: the consumer activities have the same objective considered by the producer. Nevertheless, the consumer process merges the knowledge and data layers considering the composition of different datasets, thus mapping multiple datasets to one single knowledge structure (the teleontology), instead of merging the mapping one dataset to its relative knowledge structure, as the producer process does.
  - Entity identification
  - Data mapping

The report of the work done during this phase of the methodology, has to includes also the description of the different choices made, with their strong and weak points. In other words the report should provide to the reader, a clear description of the reasoning conducted by all the different team members.

9 Evaluation

This section aims at describing the evaluation performed at the end of the whole process (producer plus consumer) over the final outcome of the iTelos methodology. More in details, this section as to report:

- the final Knowledge Graph information statistics (like, number of etypes and properties, number of entities for each etype, and so on).
- Knowledge layer evaluation: the results of the application of the evaluation metrics applied over the knowledge layer of the final KG.

- Data layer evaluation: the results of the application of the evaluation metrics applied over the data layer of the final KG.
- Query execution: the description of the competency queries executed over the final KG in order to test the suitability of the KG to satisfy the project purpose.

10 Metadata Definition

In this section the report collects the definitions of all the metadata defined for the different resources produced along the whole process (producer and consumer). The metadata defined in this phase describes both the final outcome of the project, and the intermediate outcome of each phase.

The definition of the metadata, is crucial to enable the distribution (sharing) of the resource produced. For this reason it is important to describe also where such metadata will be published to distribute the resources it describes (for example the DataScientia catalogs).

In particular the structure of this section is organized as follows, with the objective to describe the metadata relative to all the type of resources produced by the project.

- Language resources metadata description
- Knowledge resources metadata description
- Data resources metadata description

11 Open Issues

This section concludes the current document with final conclusions regarding the quality of the process and final outcome, and the description of the issues that (for lack of time or any other cause) remained open.

- Did the project respect the scheduling expected in the beginning ?
- Are the final results able to satisfy the initial Purpose ?
  - If no, or not entirely, why ? which parts of the Purpose have not been covered ?

Moreover, this section aims to summarize the most relevant issues/problems remained open along the iTelos process. The description of open issues has to provide a clear explanation about the problems, the approaches adopted while trying to solve them and, eventually, any proposed solution that has not been applied.

- which are the issues remained open at the end of the project ?