KDI : Knowledge and Data Integration

iTelos: Formal Modeling (Theory)

W7.L14.M5.T14.All

Contents

1 Data Type Definition

2 Top Level View

3 SKG Generation

4 Data Preparation

5 Phase Iterations

6 Languages and Standards

7 Tools

8 Deliverables

What are Data Types?

- In computer science, a data type or simply type is an attribute of data which tells the compiler or interpreter how the data is intended to be used.
- A data type constrains the values that an expression, such as a dataproperty (attributes) in an ontology, might take.
- Data type also defines the operations that can be done on the data, the meaning of the data, and the way values of that type can be stored.
- A data type provides a set of values from which an expression may take its values [a glimpse provided in the following two slides].

Basic Data Types

- Basic Data Types provide the basic building blocks around which Data Types are defined.
- Each Basic Data Type must include the possibility to codify the information that the value can be unknown while we do not allow the explicit storage of the null value.
- Basic Data Types (in our context) can be from any of the following three categories:-
  - **Numerical Data Types** [Boolean, Integer, Long, Float etc.]
  - **Reference Data Types** [String, NLString etc.]
  - **Semantic Data Types** [SString, Concept, Entity etc.]

Complex Data Types

- Etypes of kind Attribute are the mechanism by which Complex Data Types are defined.
- Examples of Complex Data Types can be: **Moment** and **Duration**.
  - **Moment** is encoded as an interval given by two values that specify the start point (Start) and the end point (End) in the timeline using standard date/time representation (ISO 8601). The Attribute Definitions within a Moment Complex Data Type are therefore **Start** and **End**.
  - **Duration** is encoded as two values that specify the minimum and the maximum amount of time. The Attribute Definitions within a Duration Complex Data Type are therefore **Minimum** and **Maximum**.

Contents

1 Data Type Definition

2 **Top Level View**

3 SKG Generation

4 Data Preparation

5 Phase Iterations

6 Languages and Standards

7 Tools

8 Deliverables

Top Level View

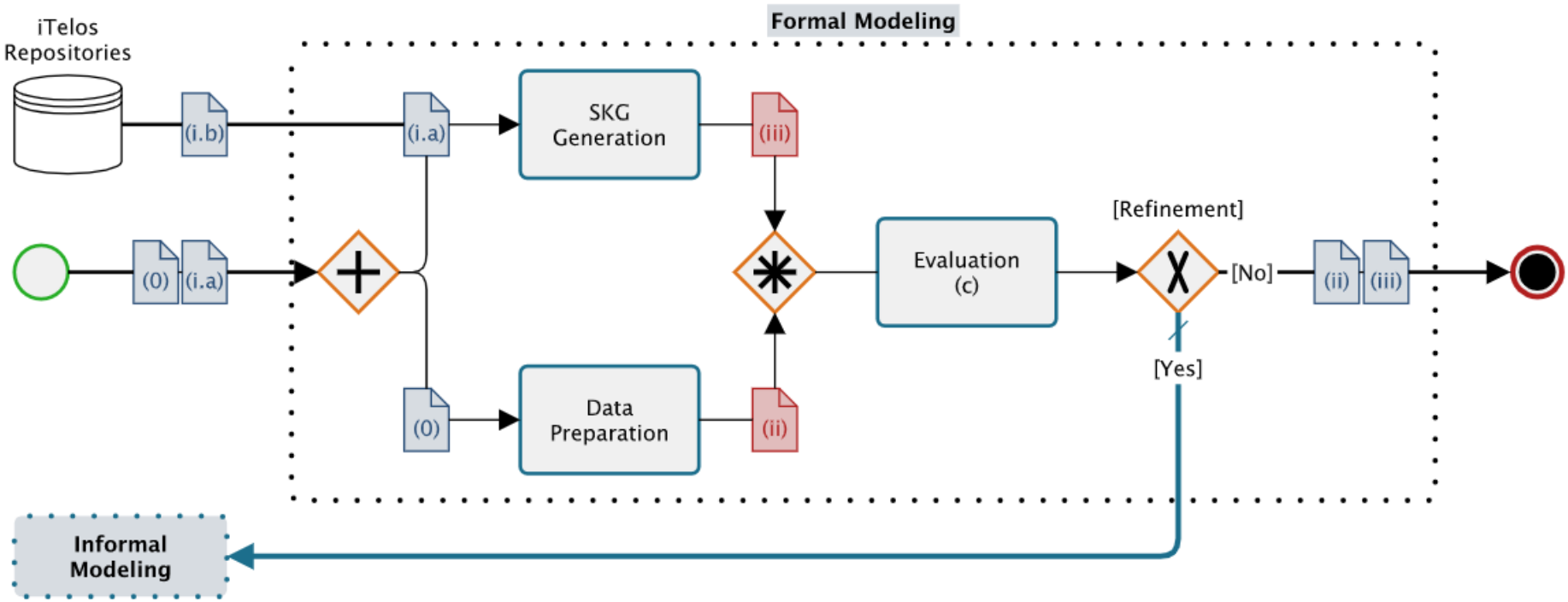
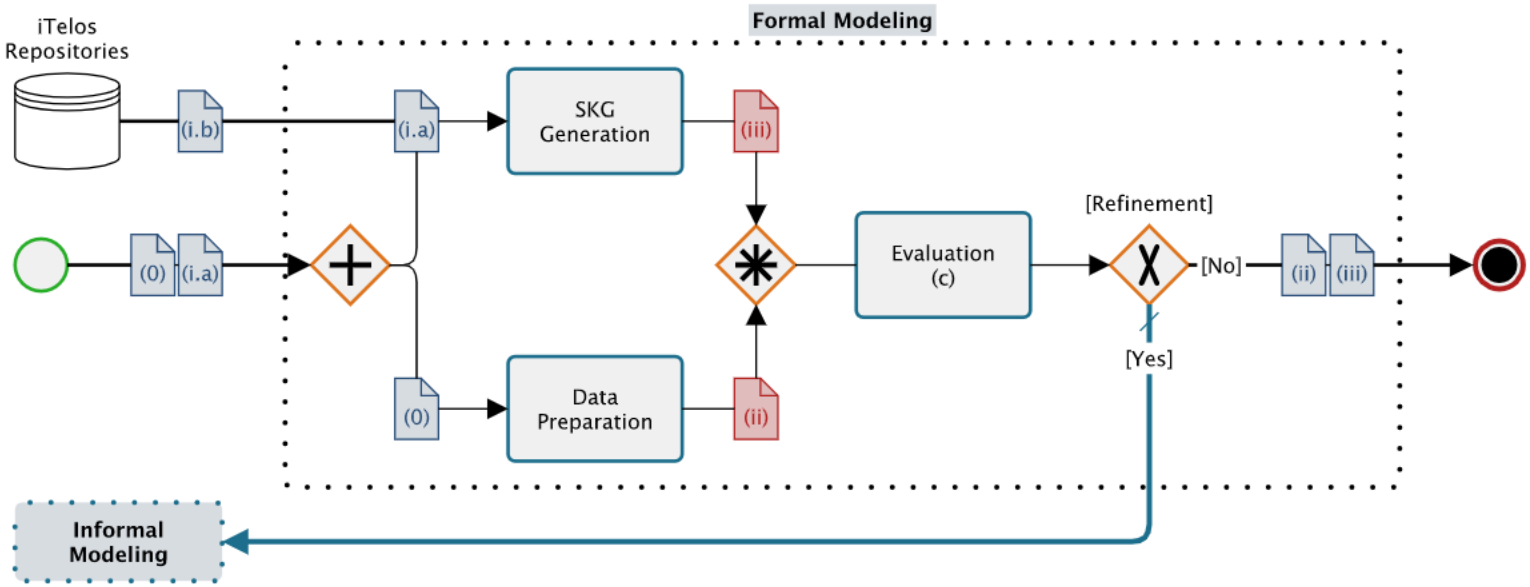


Figure: Formal Modeling Diagram

Top Level View



where:

- 0: Data sets metadata and informal metadata
- i.a: L4 informal schema
- i.b: Teleology
- ii: Aligned/Formatted data sets
- iii: L4 SKG

Top Level View

- The Formal Modeling is the fourth phase of the iTelos Methodology containing the construction of the formal L4 Schema Knowledge Graph and a first annotation regarding the L1 and L2 lexical and semantic concepts of the SKG
- The Knowledge Engineer is interested in the Schema-level macro-activities of the SKG Generation.
- The Data Scientist is interested in the Data-level macro-activities of the Data Preparation.

Top Level View

- Both roles describe and perform activities related to the advancement of the project.
- Later on, with iterations they will be combining their outputs for refining the previously produced outputs.
- The outputs of each level will be evaluated and will be either iterated for refinement (or denied for not following correctly the informal L4 schema previously made).
- If the outputs have followed the phase correctly and executed a certain amount of iterations as described per the Evaluation section, they will be accepted as formalized documents to be used in the next phase.

Contents

1 Data Type Definition

2 Top Level View

3 **SKG Generation**

4 Data Preparation

5 Phase Iterations

6 Languages and Standards

7 Tools

8 Deliverables

SKG Generation

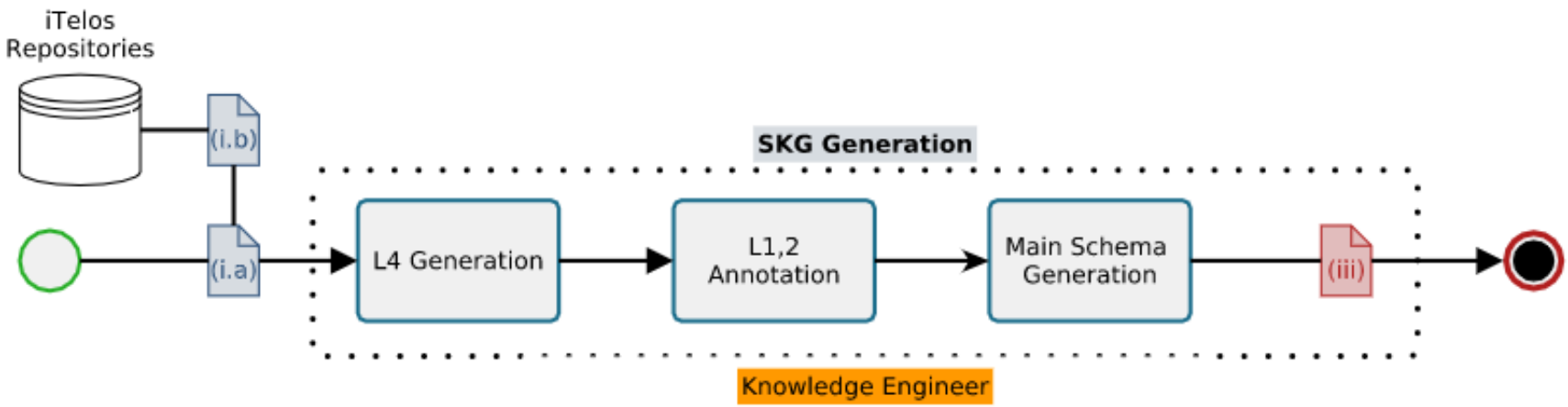
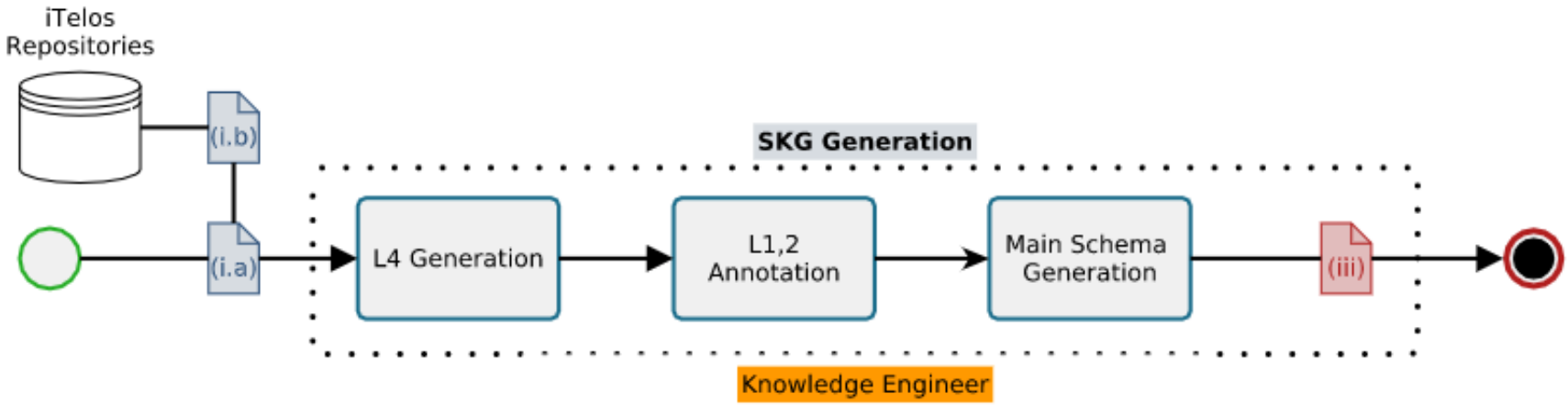


Figure: SKG Generation Diagram

SKG Generation



where the main activities being executed are the following:

- L4 Generation
- L1,2 Annotation
- Main Schema Generation

SKG Generation

- This macro-activity is found in the Schema level of the iTelos Methodology
- The Knowledge Engineer is tasked with constructing the formal L4 schema knowledge graph
- The formal L4 schema knowledge graph will be annotated through the L1 and L2 concept definitions and finally exported as the complete SKG which will be used as formal knowledge in the next section.

SKG Generation

- The main important objective is to relate the teleology found in the iTelos Repositories to help support and facilitate the construction of the L4 formal model.

- The Knowledge Engineer will later on relate the concepts as in the teleology to the entities generated for the problem and slowly categorize and export the new L1 and L2 concepts which will need to be fixed.

Teleology

A **Teleology** is constituted of three particular levels of knowledge. **L1:-**

- L1 is regarding the Concept Space
- It is interested in the linguistic interpretation of the verbs and nouns interesting to the problem.
- In here words are categorized into speech categories and are described according to their relation with other words in the system.

Teleology

A **Teleology** is constituted of three particular levels of knowledge. **L2:-**

- L2 regarding the Lexical-Semantic Space
- It is interested in the definition of the group of synsets organized according to the linguistic principles.
- L2 is usually denoted as a detailed annotation over the L1 lexicons and are described in multiple languages.

Teleology

A **Teleology** is constituted of three particular levels of knowledge. **L4:-**

- L4 regarding the representation of the SKG for a specific problem
- This schema is generated based on a data-driven requirement and uses standards already implemented (thanks to the teleologies).
- A data structure graph-like is used where the nodes define the entities and their attributes and the edges define the relations between the nodes (such as entity relation attributes).
- In this level the entities and relations are constrained by properties and logics which define theorems to proof the existence of such node or edge.

L4 Generation

- In the L4 Generation the Knowledge Engineer aims to obtain a new formal version of the knowledge schema (L4)starting from the Teleology, selected from the iTelos Repositories, and the L4 informal schema produced as output of the previous phase.
- The new L4 schema will be formally defined using RDF/OWL.
- To achieve this objective, the informal L4 schema is compared to the Teleology in order to identify in the latter, those nodes which are already able to represent the informal ETypes defined in the informal L4 schema.

L4 Generation

In case the Teleology doesn't contain enough knowledge to cover all the informal ETypes (in other words there are no nodes in the Teleology that can be used to represent some informal ETypes),

new nodes and edges are added in the new L4 formal schema (the new knowledge schema elements will be added, in the future, within the Teleology, if necessary, to produce a more defined Teleology).

L1-2 Annotation

- In the L1-2 Annotation the Knowledge Engineer has to identify within the Teleology those lexicon-semantic elements of L1 and L2, which are concepts in the UKC, that can be used to define ETypes, ETypes attributes and relations for the data which have to be integrated.
- The main purpose of the Annotation activity is to contribute into finding most, if not all, the possible concepts in the UKC that can help to describe the entities and relations found in the finalized ontology which will be used as formal knowledge definition.
- During this phase, the Knowledge Engineer, could discover within the dataset analyzed, some concepts that have not yet been added in the UKC (so not available in the Teleology)

L1-2 Annotation

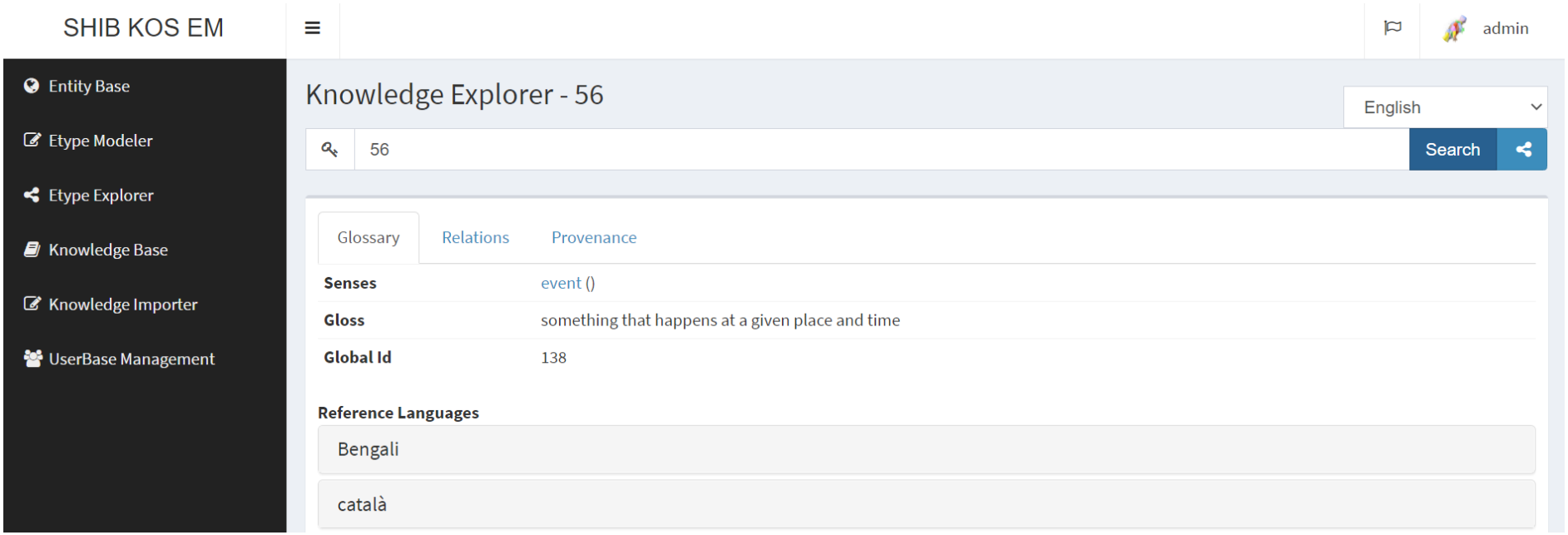
- In this case the new concepts will be added as formal knowledge and they will go to integrate the existing Teleology.
- Moreover some different representations of already existing concepts can be found within the data analyzed, so in this case those will be annotated as synonyms of the respective concepts found in the UKC.
- This sub-activity can modify the RDF/OWL file produced in the previous one, adding the L1,2 annotation, and/or produce an Import File for those new concepts/senses which have to be imported in order to integrate the UKC.
- That import file is an Excel file that lists, using a specific structure the knowledge elements to import.

L1-2 Annotation: Steps

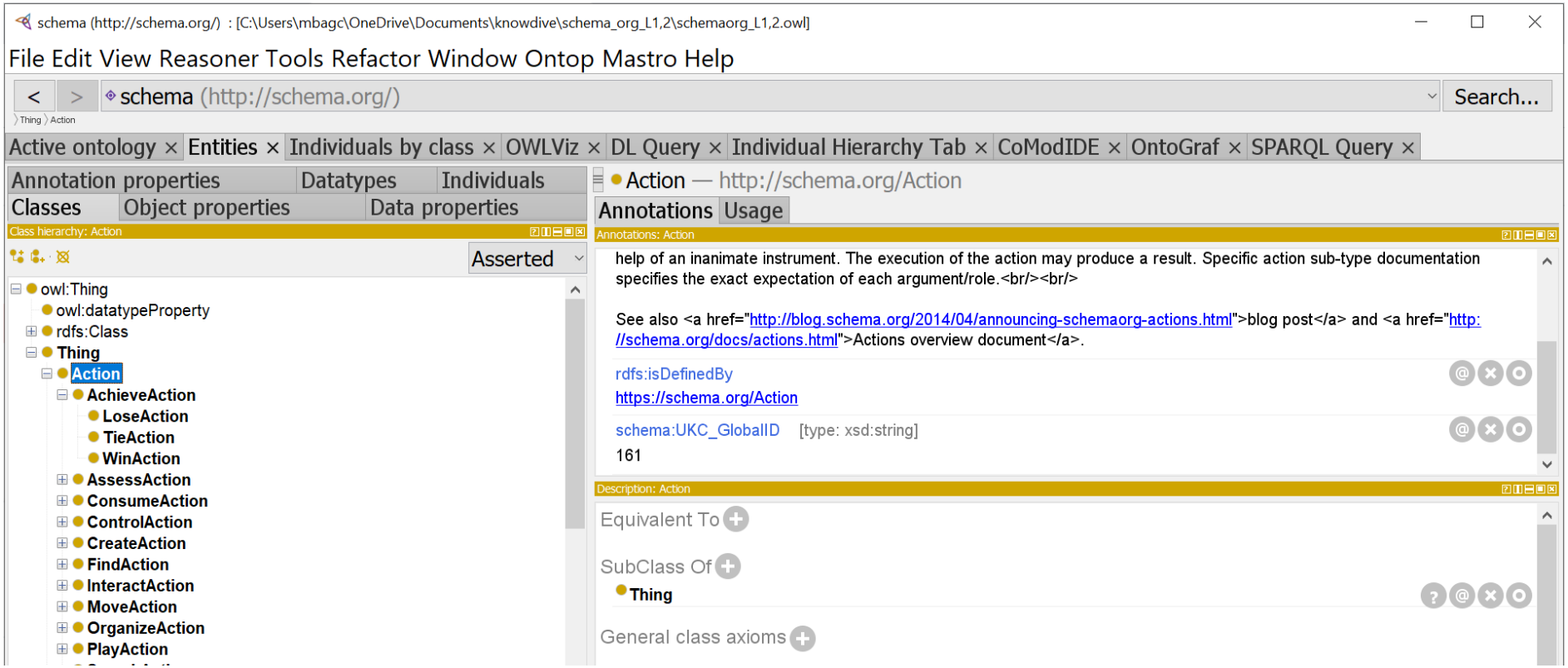
The steps of the annotation step are described as follows:

- Each individual entry term from Classes, Object Properties and Data Properties (from the L4) are selected and its meaning is intuitively understood from its description in the rdfs:comment annotation entry.
- The semantically equivalent concept of the term considered is searched for in the UKC Knowledge Base. There can be three scenarios (only one amongst three for each term):-
  - Exact Match
  - Synonym
  - Incomplete

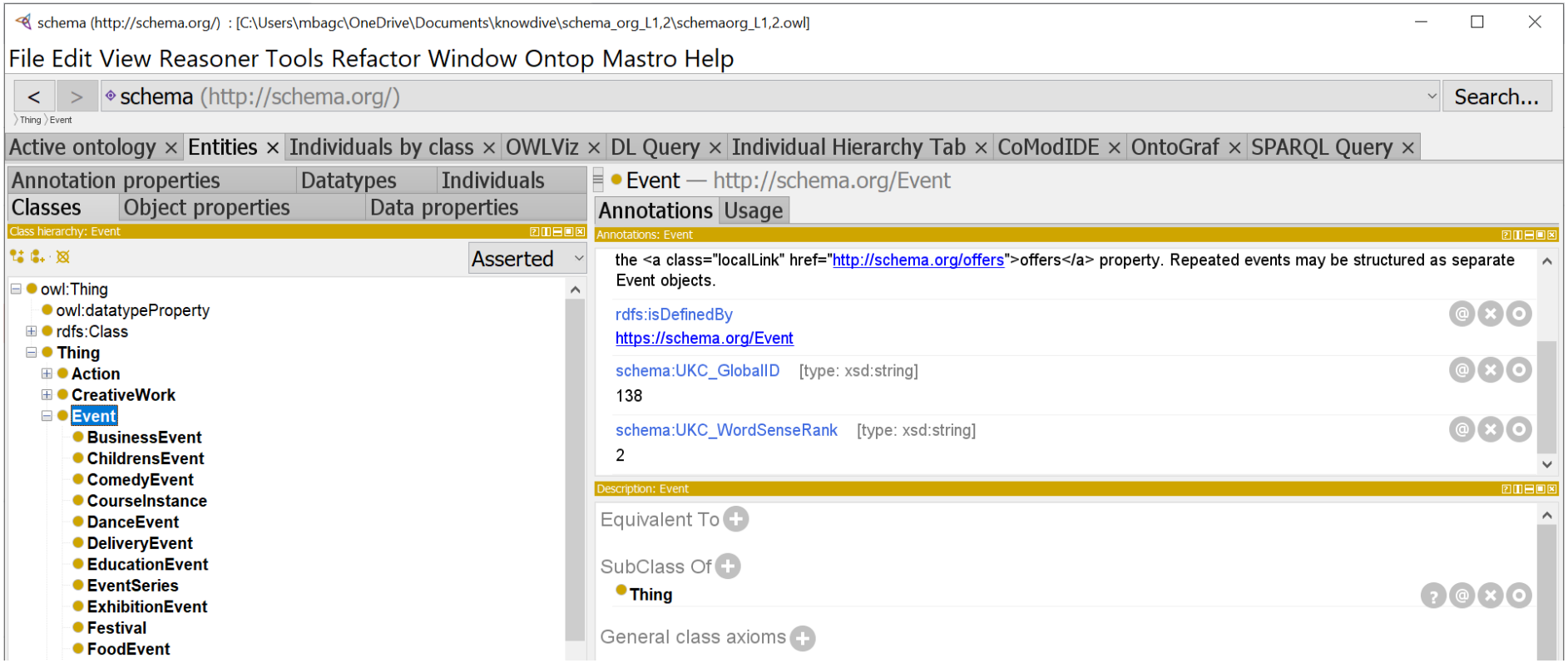
L1-2 Annotation: UKC KB Interaction



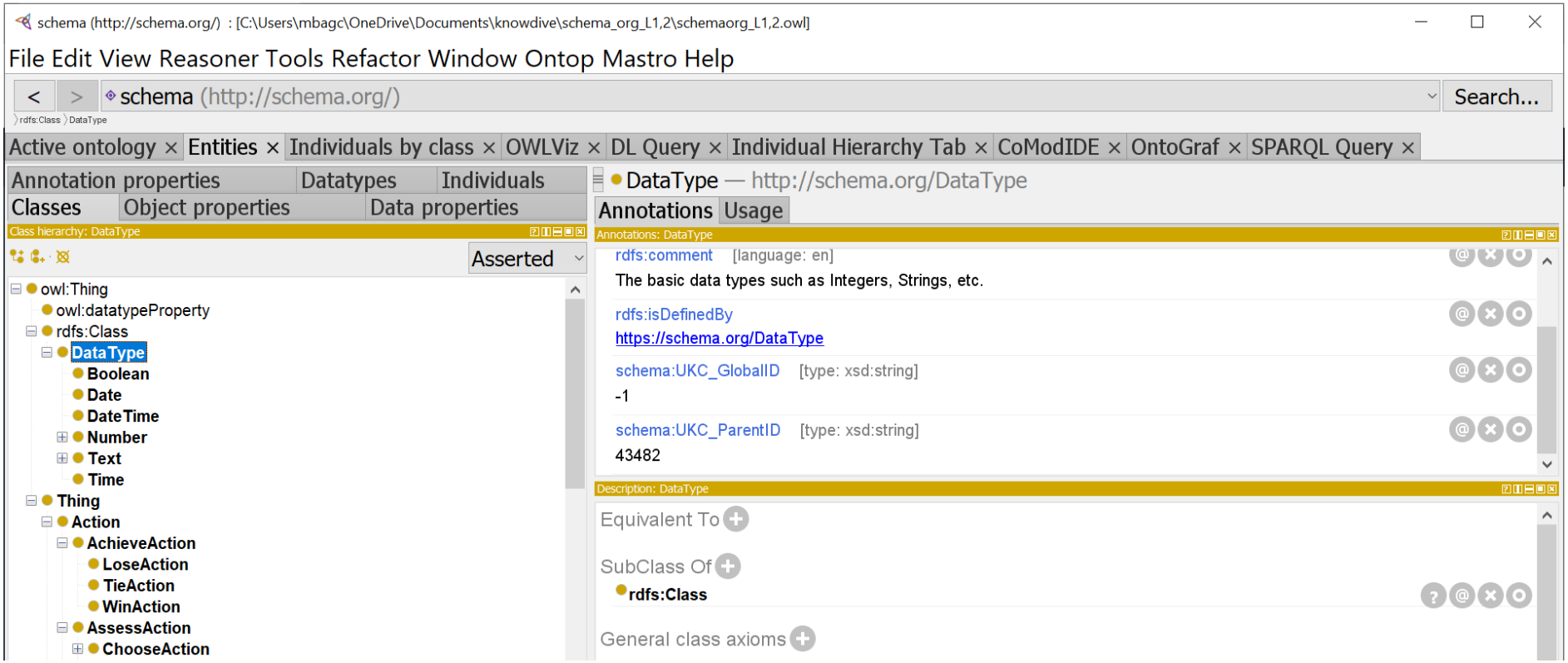
L1-2 Annotation: Exact Match



L1-2 Annotation: Synonym



L1-2 Annotation: Incomplete



Main Schema Generation

- Once the previous two sub-activities are completed, the outputs generated are the RDF/OWL file and if new knowledge elements have to be imported, the Excel import file.
- Remember to import (through the Excel file and Importer tool) all the L1-2 elements needed **before** generating the final SKG.
- In the current sub-activity these two objects are imported together, using the appropriate tools, in the Data Integration Platform where they will be merged in a single, and more precise, SKG that will be exported and provided as the output of this sub-activity.

Contents

1 Data Type Definition

2 Top Level View

3 SKG Generation

4 **Data Preparation**

5 Phase Iterations

6 Languages and Standards

7 Tools

8 Deliverables

Data Preparation

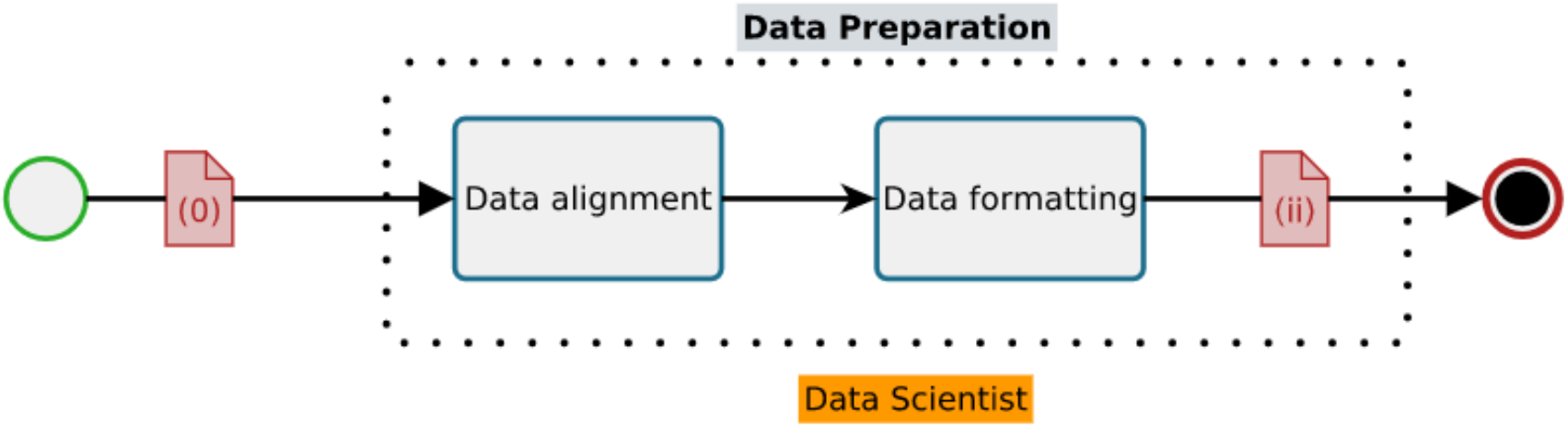
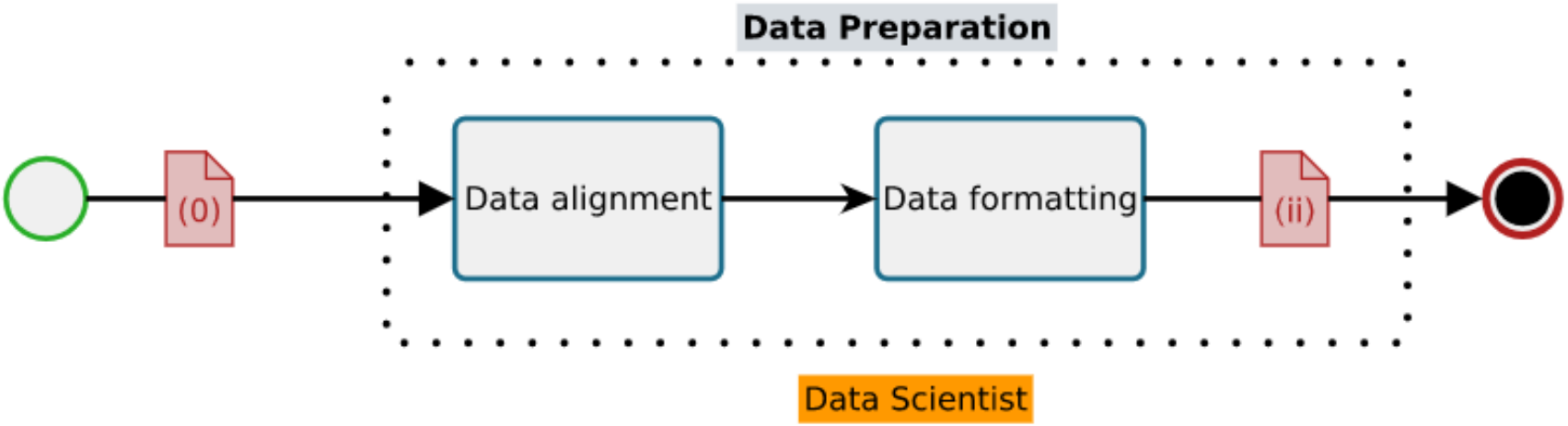


Figure: Data Preparation Diagram

Data Preparation



where the main activities being executed are the following:

- Data alignment
- Data formatting

Data Preparation

- The Data Preparation macro-activity aims to handle the data identified and extracted during the previous phases(Inception and Informal modelling), and perform shaping operations in order to provide them to the next phase in the best form as possible.

- These operations regard two different aspects, the data alignment respect the ETypeused to represent the data, and the data formatting following the correct data types, the two sub-activities manage these actions respectively.

Data alignment

- The Data alignment activity compares the data extracted from the Informal Modelling phase, with the informal definition of the ETypes, trying to understand if the data are correctly shaped to be represented by those ETypes.

- In case differences appear between the data form and the ETypes structure, this activity aims to reorganize the data with the objective to reduce as much as possible the gap between the data layer and knowledge layer.

Data formatting

- The Data formatting activity has the same objective of the previous activity, but focused on the values of the data instead of their structure.

- It checks and models the data values to ensure that those values respect the data types for the data that they represent.

- The information about the correct data types to adopt comes from the knowledge defined for the respective data.

- This activity aims to identify and, if necessary provide solutions to those data values that are no type compliant.

Contents

1 Data Type Definition

2 Top Level View

3 SKG Generation

4 Data Preparation

5 **Phase Iterations**

6 Languages and Standards

7 Tools

8 Deliverables

Phase Iterations

- In the Formal Modeling phase the bare minimum iterations required for the production of an high quality output is expected after four iterations.

- The iterative process in this phase is scheduled in order to assign one main type of data (Core, Common and Contextual) to each iteration, plus one final iteration for a final check.

- Each iterations aim to improve step by step, the ontology that will be used in the next phase as formal knowledge of the data to be integrated, for the schema level, and for the data level, a set of data as much as possible aligned to the ontology and well shaped in terms of values and data types.

Iteration Zero

- In the first iteration the main output of the Schema level is the definition of formal ETypes for the Core entities,starting from the Informal definition provided by the previous phase. As part of the formal definition the EType are annotated with the lexicon-semantic element of L1 and L2.

- During the first iteration the Data level is less considered due to the fact that the formal definition of the ETypes is needed to perform the data preparation sub-activities. However a first general iteration of Data Alignment and Data Formatting activities, can be done on the base of ETypes informal definitions coming from the previous phase.

Iteration One

- In the second iteration, in the Schema level, the ontology that includes the formal definition of the Core ETypes,is improved adding the formal definition of the Common ETypes annotated with the lexicon-semantic element L1 and L2.

- In the Data level the Core entity data are handled, following the two sub-activities of the Data Preparation macro-activity, on the base of the Core formal ETypes defined at Schema level in the previous iteration.

Iteration Two

- In the third iteration, in the Schema level, the ontology that includes the formal definition of the Core and Common ETypes, is improved adding the formal definition of the missing Contextual ETypes annotated with the lexicon-semantic element L1 and L2.

- In the Data level the Common entity data are handled, following the two sub-activities of the Data Preparation macro-activity, on the base of the Common formal ETypes defined at Schema level in the previous iteration.

Iteration Three

- In the fourth iteration, in the Schema level, the ontology includes the formal definition of the Core, Common and Contextual ETypes for the data that have to be integrated, for this reason this iteration is used as a final check inorder to identify missing knowledge definitions.

- In the Data level the missing Contextual entity data are handled, following the two sub-activities of the Data Preparation macro-activity, on the base of the Contextual formal ETypes defined at Schema level in the previous iteration.

- After this minimum number of iteration the result of this phase is produced and can be moved on to the next phase, i.e. the ontology containing the knowledge formal definition for the dataset and the dataset aligned to the knowledge and well shaped.

Contents

1 Data Type Definition

2 Top Level View

3 SKG Generation

4 Data Preparation

5 Phase Iterations

6 **Languages and Standards**

7 Tools

8 Deliverables

Languages and Standards

- During the Formal modeling the Knowledge Engineer define formally the EType and their relations producing the L4 schema annotated with the L1-2 elements coming from the UKC. The L4 schema is defined using the RDF-OWLformat. For this reason the knowledge of that standard is required in the current phase.

- Moreover the Knowledge Engineer has to be able to import new L1-2 knowledge elements, if they are missing, in order to correctly define the data used in the project. To do that she/he has to know the KnowDive internal standard used to define the Excel Import file (L1-2 Import File standard).

Contents

1 Data Type Definition

2 Top Level View

3 SKG Generation

4 Data Preparation

5 Phase Iterations

6 Languages and Standards

7 **Tools**

8 Deliverables

Tools

In the Formal Modeling phase the Knowledge Engineer have to define the Schema Knowledge Graph (SKG). In order to do that the tool used are:

- **Protégé:** This tool is used to define the SKG using the OWL format.
- **UKC:** The usage of the UKC is provided to the Knowledge Engineer, in order to annotate the L4 schema, previously generated with the already existing L1-2 elements.
- **Knowledge Importer:** Thanks to this tool, the Knowledge Engineer can import new L1-2 elements (new concepts and relations) in the UKC, and so let them available to annotate the SKG that has to be produced.

Tools

In the Formal Modeling phase the Knowledge Engineer has to define the Schema Knowledge Graph (SKG). In order to do that the tool used are:

- **OWL importer:** This tool allows the import of the whole SKG within the Data Integration platform, in order to be able to improve the knowledge layer of an eventually existing SKG within the platform itself.
- **OWL exporter:** This tool allows to export, from the Data Integration platform the OWL file which defines the whole SKG contained in the platform. This OWL file can be used as knowledge input for the KarmaLinker tool in the next Data Integration phase.
- At data level the Data Scientist can use the Jupyter environment, previously set up, to align and format the data, providing so well formed dataset to the next final phase.

Contents

1 Data Type Definition

2 Top Level View

3 SKG Generation

4 Data Preparation

5 Phase Iterations

6 Languages and Standards

7 Tools

8 **Deliverables**

Deliverables

In the Formal Modelling phase there aren't new deliverable produced, but some crucial improvements are done on the documents already existing

- **iTelos project report:** The main project report is updated with the formal definition of the ETypes, and most important, the definition of the L4 Schema produced in output in the current phase, together with the description of the Schema annotation using the L1-2 elements.
- **Metadata sheet and description:** The metadata documents are updated extending both the set of metadata collected and the relative description in the description document. In the current phase, these documents are updated with the L1-2 and L4 metadata information.

KDI : Knowledge and Data Integration

W7.L14.M5.T14.All

iTelos: Formal Modeling (Theory)