KDI Knowledge and Data Integration

iTelos - Inception

W3.L6.M3.T7

Contents

1. Top level view
2. Schema Inception
3. Data Inception
4. Evaluation
5. Phase Iterations
6. Languages & Standards
7. Tools
8. Deliverables
9. Examples

Contents

1. **Top level view**
2. Schema Inception
3. Data Inception
4. Evaluation
5. Phase Iterations
6. Languages & Standards
7. Tools
8. Deliverables
9. Examples

Top level view

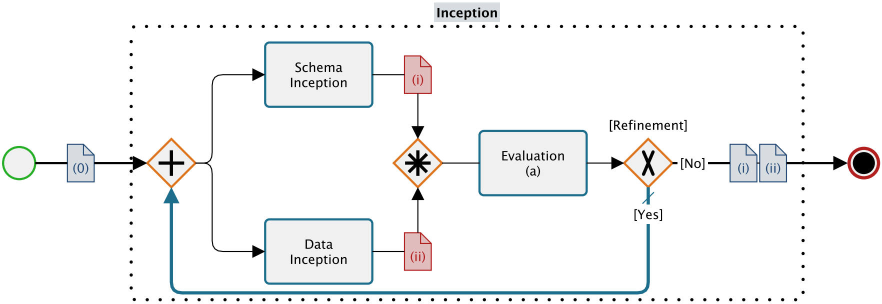
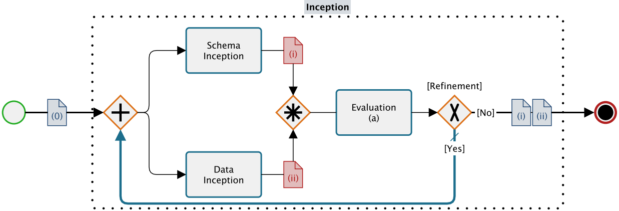


Figure: Inception Diagram

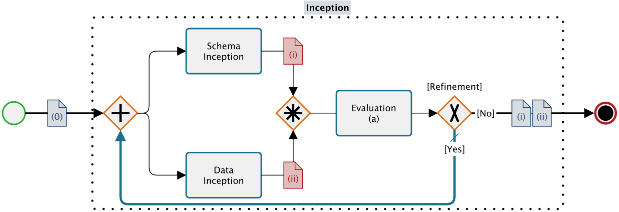
Top level view



where:

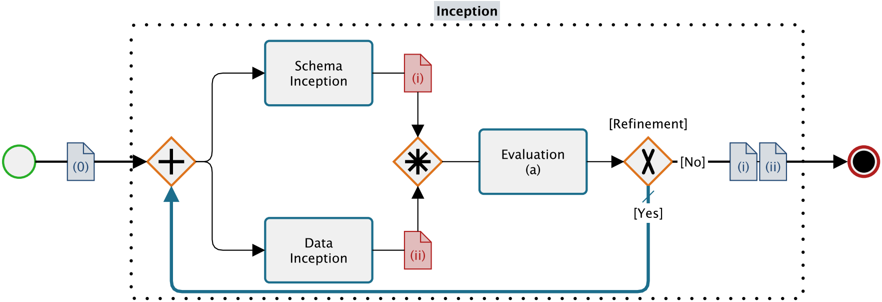
- 0 : Purpose Documentation.
- i : Competency Queries with Data Objects definition.
- ii : Preliminary datasets and informal metadata.

Top level view



The *Inception* phase aims to define what are called *Competency Questions* (**cq**), that at the end of this phase will became *Competency Queries* defining all kinds of queries that can be generated to solve the problem.

Top level view

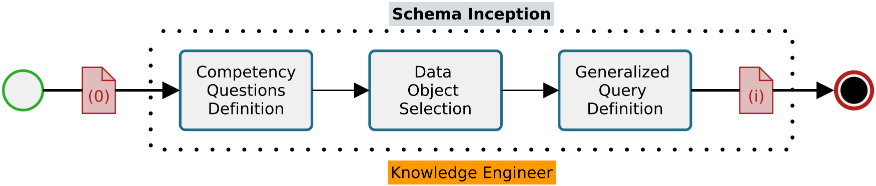


The **Knowledge Engineer** and the **Data Scientist** are respectively in charge of the **Schema Inception** and **Data Inception** activities.

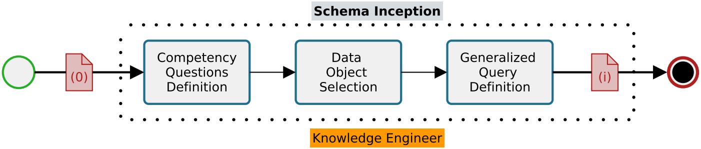
Contents

1. Top level view
2. **Schema Inception**
3. Data Inception
4. Evaluation
5. Phase Iterations
6. Languages & Standards
7. Tools
8. Deliverables
9. Examples

Schema Inception



Schema Inception



In the Schema Inception the main sub-activities being executed are the following:

- Competency Questions Definition.
- Data Object Selection.
- Generalized Query Definition.

Competency Questions Definition - 1/2

In this first sub-activity the Knowledge Engineer using the Purpose documentation, has to define the **cq.**

The Knowledge Engineer categorizes the **cq** collected during the phase iterations, following the different data typologies:

- Core data**.**
- Common data**.**
- Contextual data**.**

Competency Questions Definition - 2/2

It is important to note that:

- The three data typologies listed above, define a *dependency hierarchy*.
- The Common data have the strongest impact in terms of dependencies.
- The Core data are the most important entities regard the project’s solution.
- The Contextual data are a specification of the previous typologies of data.

Data Object Selection

Data Object Selection starts using the **cq** definitions.

The scope of this internal step is to identify and list the main data object involved in the questions defined before.

This general object are the first version of what will be called **etype**, and in the current phase are used in the next sub-activity to improve the **cq** definition.

Generalized Query Definition

This sub-activity aims to define more precisely all kinds of queries which can be useful in the solution achievement.

To obtain this result, the Knowledge Engineer:

- uses the preliminary defined **cq** together with the data object defined in the previous step,
- proceed to write a list of queries in a more precise format (i.e. SQL-like language).

Contents

1 Top level view

2 Schema Inception

3 **Data Inception**

4 Evaluation

5 Phase Iterations

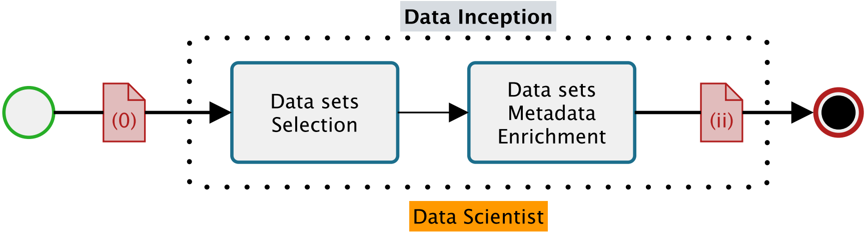
6 Languages & Standards

7 Tools

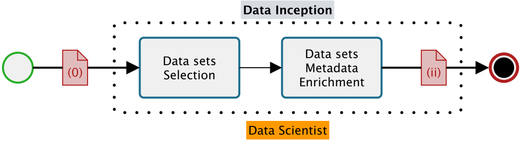
8 Deliverables

9 Examples

Data Inception



Data Inception



In the Data Inception the sub-activities being executed are the following:

- Data sets Selection.
- Data sets Metadata Enrichment.

Data sets Selection

- The Data Scientist, has to analyze all the data sources listed in the Purpose documentation.
- The Data Scientist (working in parallel with the Knowledge Engineer) has to identify the correct datasets respect to the users and scenarios previously defined.
- Some data sources could be not a repository but instead specific web location from where a scraping procedure is needed in order to extract the data.

Data sets Metadata Enrichment

- The Data Scientist is tasked on enriching the datasets selected and extracted with record-level metadata.

- Part of the metadata might have to be discovered by reading the importance of the problem context.

Contents

1 Top level view

2 Schema Inception

3 Data Inception

4 **Evaluation**

5 Phase Iterations

6 Languages & Standards

7 Tools

8 Deliverables

9 Examples

Evaluation - 1/4

- The main aspects for Inception evaluation is the alignment between project informal knowledge collected and datasets collected.

- The output of this phase is made out of **cq**s plus the preliminary list of data sets with informal **metadata** selected out of the problem documentation.

- The class and properties are collected into two types of sets, S_c and S_p from the **cq**s and the reference alignment data set.

Evaluation - 2/4

Definition

Coverage (Cov) is the coverage between two sets α onto β, the percentage of the difference from α from β.

$$Cov(\alpha, \beta) = 1 - \frac{|\alpha - \beta|}{|\alpha|}$$

Definition

Flexibility (Flx) is the flexibility between two sets α onto β, the percentage of the difference from α from β.

$$Flx(\alpha, \beta) = \frac{|\beta - \alpha|}{|\beta|}$$

Evaluation - 3/4

The Coverage calculates the ratio $\alpha \cap \beta$ to α which is the percentage of the join set to the source set α.

$Cov(C_{c/p}, D_{c/p})$ evaluates the percentage of the overlapped part of **cq**s, where C and D stand for **cq** and the referenced alignment data set.

c/p stands for the type of set, classes as c and properties as p.

Evaluation - 4/4

The Flexibility returns the ration $\beta - \alpha$ to β which is the percentage of the leftover of the target set β to itself.

$Flx(C_{c/p}, D_{c/p})$ evaluates the leftover percentage of the reference alignment data sets.

Contents

1. Top level view
2. Schema Inception
3. Data Inception
4. Evaluation
5. **Phase Iterations**
6. Languages & Standards
7. Tools
8. Deliverables
9. Examples

Phase Iterations

In the Inception **phase** the minimum number of **iteration**s required for the production of an high quality output is equal, or more than, four iterations.

The iterative process provide at each iterations a more defined and precise input to the activities.

More in detail each iteration is focused on a specific data typology, among the three already mentioned, Common, Core and Contextual.

Iteration Zero

In the first iteration the main output of the schema level is a document with the definitions of general queries with the general object definition for the Common data typology.

Regarding the data level, instead, the first iteration aims to identify eventually missing data sources.

Iteration One

The Knowledge Engineer has to define the general queries and the general objects definition for the Core data typology.

At data level the Data Scientist has to extract the dataset and collect metadata for the Common typology.

The evaluation activity at the and of the iteration, verifies the datasets extracted using the schema elements defined in the parallel activity

Iteration Two

The Knowledge Engineer has to define the general queries and the general objects definition for the Contextual data typology.

At data level the Data Scientist has to extract the dataset and collect metadata for the Core typology.

The evaluation activity at the and of the iteration, verifies the datasets extracted using the schema elements defined in the parallel activity.

Iteration Three

The Knowledge Engineer has to perform a general check on the documentation produced in the previous iteration.

At data level the Data Scientist has to extract the dataset and collect metadata for the Contextual typology.

Contents

1 Top level view

2 Schema Inception

3 Data Inception

4 Evaluation

5 Phase Iterations

6 Languages & Standards

7 Tools

8 Deliverables

9 Examples

Languages & Standards

In this phase the Knowledge Engineer has to produce as final schema level output the Generalized Query Definitions.

The Data Scientist starts to collect the dataset needed from the data sources identified in the previous phase.

The datasets collected can be expressed using one or more of the following format:

- XML
- HTML
- CSV
- JSON

Contents

1. Top level view
2. Schema Inception
3. Data Inception
4. Evaluation
5. Phase Iterations
6. Languages & Standards
7. **Tools**
8. Deliverables
9. Examples

Tools

The Knowledge Engineer can use a spreadsheet tool such as Excel or Google Sheet to produce the documentation containing initially the **cq**s.

For the data level the Scientist has to collect and manage data from the data sources defined in the previous phase.

Examples of libraries that can be used:

- **Data management**: Pandas, NumPy, Scikit Learn
- **Data scraping**: Beautiful Soup, Scrapy, Selenium, LXML
- **Data formatting**: Arrow, PrettyPandas, datacleaner

Contents

1 Top level view

2 Schema Inception

3 Data Inception

4 Evaluation

5 Phase Iterations

6 Languages & Standards

7 Tools

8 **Deliverables**

9 Examples

Deliverables - 1/2

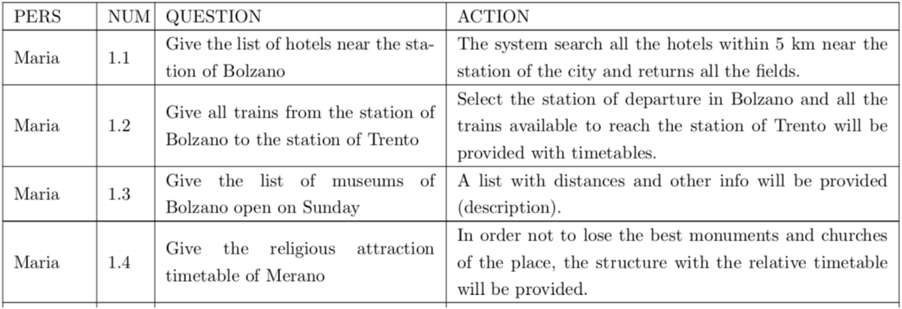
In the Inception phase the main deliverables produced are:

- **iTelos project report.**
- **Preliminary datasets sheet.**
- **Metadata sheet.**
- **Metadata description.**

Contents

1. Top level view
2. Schema Inception
3. Data Inception
4. Evaluation
5. Phase Iterations
6. Languages & Standards
7. Tools
8. Deliverables
9. **Examples**

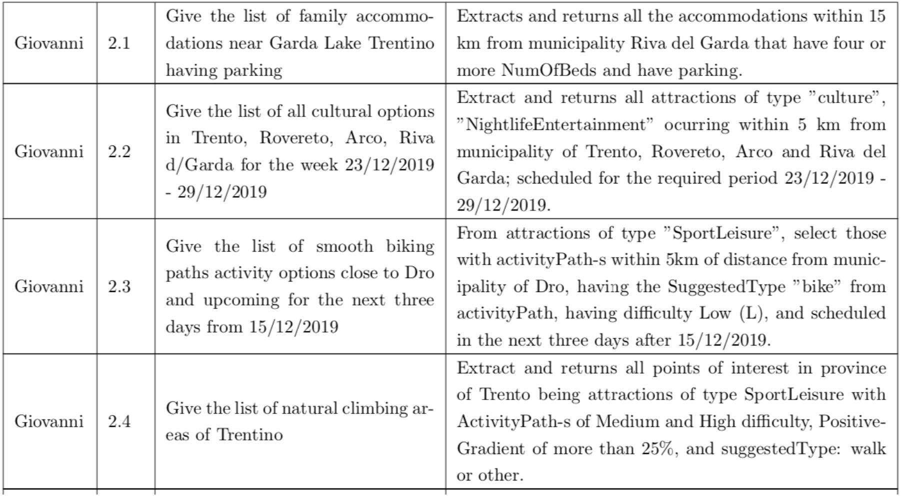
Examples of Schema Inception



PERS	NUM	QUESTION	ACTION
Maria	1.1	Give the list of hotels near the station of Bolzano	The system search all the hotels within 5 km near the station of the city and returns all the fields.
Maria	1.2	Give all trains from the station of Bolzano to the station of Trento	Select the station of departure in Bolzano and all the trains available to reach the station of Trento will be provided with timetables.
Maria	1.3	Give the list of museums of Bolzano open on Sunday	A list with distances and other info will be provided (description).
Maria	1.4	Give the religious attraction timetable of Merano	In order not to lose the best monuments and churches of the place, the structure with the relative timetable will be provided.

Figure: Space Domain Competency Questions

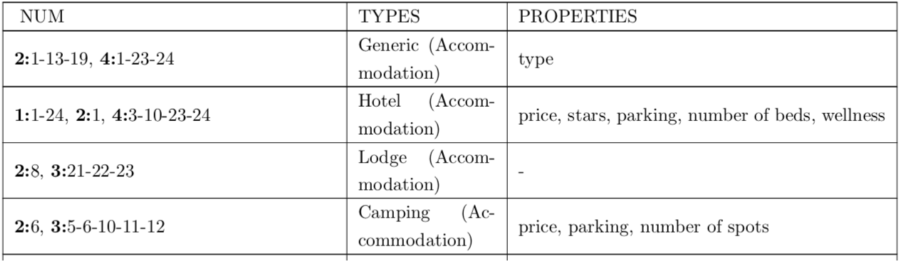
Examples of Schema Inception



Giovanni	2.1	Give the list of family accommodations near Garda Lake Trentino having parking	Extracts and returns all the accommodations within 15 km from municipality Riva del Garda that have four or more NumOfBeds and have parking.
Giovanni	2.2	Give the list of all cultural options in Trento, Rovereto, Arco, Riva d/Garda for the week 23/12/2019 - 29/12/2019	Extract and returns all attractions of type "culture", "NightlifeEntertainment" ocurring within 5 km from municipality of Trento, Rovereto, Arco and Riva del Garda; scheduled for the required period 23/12/2019 - 29/12/2019.
Giovanni	2.3	Give the list of smooth biking paths activity options close to Dro and upcoming for the next three days from 15/12/2019	From attractions of type "SportLeisure", select those with activityPath-s within 5km of distance from municipality of Dro, having the SuggestedType "bike" from activityPath, having difficulty Low (L), and scheduled in the next three days after 15/12/2019.
Giovanni	2.4	Give the list of natural climbing areas of Trentino	Extract and returns all points of interest in province of Trento being attractions of type SportLeisure with ActivityPath-s of Medium and High difficulty, PositiveGradient of more than 25%, and suggestedType: walk or other.

Figure: Space Domain Competency Questions

Examples of Schema Inception



NUM	TYPES	PROPERTIES
2:1-13-19, **4:**1-23-24	Generic (Accommodation)	type
1:1-24, **2:**1, **4:**3-10-23-24	Hotel (Accommodation)	price, stars, parking, number of beds, wellness
2:8, **3:**21-22-23	Lodge (Accommodation)	-
2:6, **3:**5-6-10-11-12	Camping (Accommodation)	price, parking, number of spots

Figure: Space Domain Query Patterns

KDI : Knowledge and Data Integration

- W3.L6.M3.T7
- iTelos - Inception