KDI : Knowledge and Data Integration

Purpose Formalization & Resource Classification

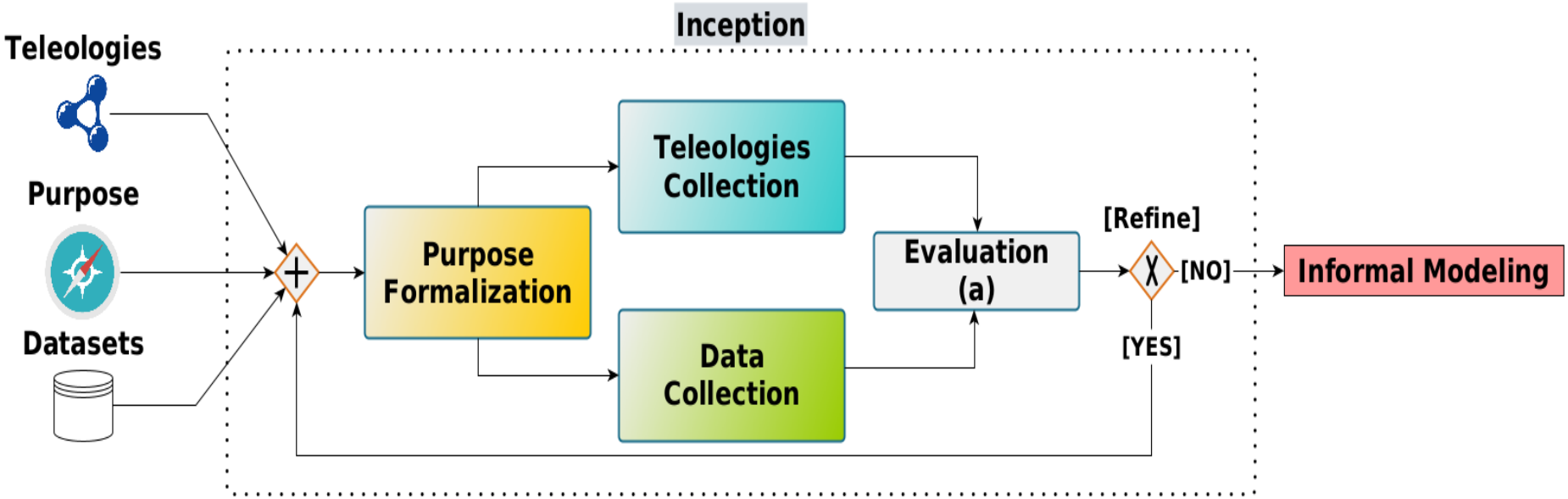
iTelos Inception Phase

Simone Bocca

Contents

Contents

Inception phase



Inception phase's objectives

Inception is the first iTelos phase

Inputs:

- Purpose.
- List of data sources.
- List of reference teleologies.

Outputs:

- Classified Competency Question (CQ) list.
- Datasets.
- Reference teleologies.

The Inception phase has two main objectives:

- start the **Purpose formalization process** that will be carried on during the second and third phase, leading to the ETG generation.
- **collect** the data and knowledge resources to be integrated and **classify** their elements into the reusability categories.

Contents

Reusability Categories

In order to identify the level of reusability of the resources collected and handled along the methodology, all the resource elements (such as entities, entity properties, ETypes, and others) are classified into one of three reusability categories, defined as follow:

- **Common**: this category involves resources associated to aspects which are common to all domains, also outside the project's DoI.
- **Core**: this category involves resources associated to the more core aspects of DoI. They carry information about the most important aspects considered by the purpose, information without which it would be impossible to develop the EG.
- **Contextual**: this last category involves resources which carry specific, possibly unique, information from the domain of interest. These are the resources whose main goal is to create added value. If core resources are necessary for a meaningful application, contextual resources are the ones which can make the difference with respect to the competitors.

Reusability Categories - Example

Let's consider a portion of the Health domain which includes all the health facilities in Trento (Italy) and all the medical care that they can offer to the citizens.

Consider now the following project's Purpose:

> *A person living in Trento, wants to easily access the best medical care she/he needs, declaring specific symptoms*

Information element examples classified:

- **Common**: *Location*, *Public Transport*, *Person*, ...
- **Core**: *Hospital*, *Pharmacy*, *Drug*, *Patient*, *Doctor*, ...
- **Contextual**: *Drug's component*, *Medical Prescription*, *Hospital Department*, *Doctor Specialization*, *Patient Allergies*, ...

Contents

Domain, Personas & Scenarios

The very first step required in the Inception phase is the definition of the context considered in the DI project. Such a context is described using three elements:

- **Domain of Interest** (DoI): the domain considered for the project. Describing the DoI, it is important to define the boundaries within which the project's information is considered.
- **Personas**: examples of all the possible actors, living in DoI, who have interest in exploiting the DI project results.
- **Scenarios**: examples of possible use cases in which the actors, previously described, can be involved.

The DoI definition as well as the description of Personas and Scenarios, allows a better understanding of the user's Purpose that starts to be formalized in the Inception phase.

Competency Questions (CQs)

- The Competency Questions (CQs) are, concretely, natural language questions defined by the DE, representing the final user needs.
- Such questions are produced considering the initial project's Purpose and, due to that, they can be seen as a first level (informal) formalization of the Purpose itself.

> The CQs are, by definition, the set of functional user requirements for the whole DI project. For this reason they will be used for the evaluation of the final output, testing the EG.

In the first (Inception) and second (Informal Modeling) iTelos phases, the CQs are submitted to a **formalization process**, which aims to extract from, the initial natural language question, a conceptualization of the user requirements (what will be called the ETG).

CQs formalization process

The CQs formalization process is composed by 5 steps executed in the first two iTelos phases:

Inception phase:

1. **Raw CQ** (= informal CQ as elicited from purpose/objectives)
2. **Kernel CQ** (= Raw CQ minus all the auxiliary/apparatus words, resulting in each term denoting a concept)
3. **Analysed CQ** (= each concept in Kernel CQ classified as common, core or contextual concepts)

Informal Modeling phase

4. **Classified CQ** (= each concept in Analysed CQ further classified as Objects, Functions and Actions)
5. **Attributed CQ** (= each concept in Classified CQ used to represent ETypes, object properties and data properties)

CQs formalization - Row CQ

We take a small motivating example from the *the domain of facilities for food and accommodation in Trentino* to illustrate our methodology-

- **1. Raw CQ** (= informal CQ as elicited from purpose/objectives)

  1. *"What are all the MS student hostels in Trentino having a consierge?"*

  2. *"List all the PhD students residing in Opera Universitaria"*

CQs formalization - Kernel CQ

- **2. Kernel CQ** (= Raw CQ minus all the auxiliary/apparatus words, resulting in each term denoting a concept)
  - *For (1): Person, MS Student, Establishment, Hostel, Consierge*
  - *For (2): Person, PhD Student, Establishment, Hostel*

CQs formalization - Analysed CQ

- **3. Analysed CQ** (= each concept in Kernel CQ classified as common, core or contextual concepts)

1. *For (1):*
   1. **COMMON:** Person, Establishment
   2. **CORE:** Hostel, Consierge
   3. **CONTEXTUAL:** MS Student

2. *For (2):*
   1. **COMMON:** Person, Establishment
   2. **CORE:** Hostel
   3. **CONTEXTUAL:** PhD Student

Contents

Resource Collection

- The second main objective of the Inception phase is to collect the data, and knowledge, resources to be integrated.
- This task can involve the usage of scripts, code libraries and/or specific tools in order to automate the resource collection, if the amount of resources considered become too large.
- The resources collected have to be properly described. During such a task, based on the project's Purpose and the Analysed CQs, the data and knowledge resources elements are classified into one of the three reusability categories (Common, Core and Contextual).

Note: see slides on *Data Collection and Preparation* for more details about the iTelos input resources.

Notes on classification

- The DE expert performing the first 3 steps of the CQs formalization process doesn't need any data management and/or knowledge representation skills.
- The classification of the resources involved in the DI projects gives important information about the resources considered:
  - **Project support level**: in order to develop the DI project and achieve a satisfactory result for the purpose, a proper number of common, core and contextual information are needed. If the set of data and/or knowledge resource doesn't respect such criteria a weak final result will be achieved (the Inception evaluation activity is in charge to check this aspects).
  - **Information availability**: a datasets involving a lot of contextual elements indicates an high level of information detail exploitable for the project's Purpose.
  - **Resource values**: often the contextual resource being very specific, are also those more difficult to retrieve and expansive in terms of costs. Based on that we can associate to each resource a value with respect to their level of reusability.

Contents

Purpose formalization & classification Summary

In this lecture we discussed:

- The phase objectives.
- The reusability categories.
- How to produce and formalize a list of CQs starting from the project's purpose.
  - How to classify the main concepts extracted from the CQs
- How to classify the resources collected for the DI project.
- Why the reusability classification is crucial in iTelos.

KDI : Knowledge and Data Integration

Simone Bocca

Purpose Formalization & Resource Classification

iTelos Inception Phase