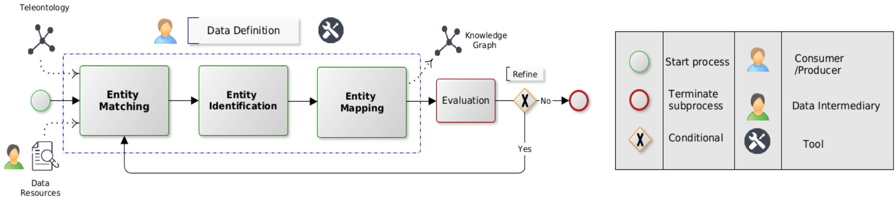
Part 5.7
Phase 5 - Entity Definition

Phase 5 - Entity Definition



- **Input**: the data resources cleaned and aligned, plus the teleontology(ETG).
- **Output**: the final KG(s).
- **Objective**: the last phase of the methodology aims at merging the knowledge and the data layers into a single structure.

Entity Definition - Objective

- To recap, in the previous phases we handled:
  - the **sources heterogeneity**, by selecting the trusted data sources;
  - the **format heterogeneity**, by formatting the resources collected adopting well-known reference open standards, and language concepts;
  - the **structure heterogeneity**, by defining a purpose-specific reusable teleontology, reusing reference standard ontologies.

Entity Definition - Objective (2)

- We start the last iTelos phase by having the most formalized version of the initial purpose:
  - The Teleontology
- Nevertheless, the teleontology defines an homogeneous representation of the information to be used to satisfy the purpose,
- **but**, it doesn't consider the **data values heterogeneity** present in the **data (values)** to be associated to the teleonotolgy.

Entity Definition - Data Values Heterogeneity

- Even fixing a source of information from which data is collected and represented through a specific data formats, as well as adopting clear data structures, a final layer of heterogeneity has to be considered.
- **Data Values Heterogeneity**, is defined over the values of the information properties which can be used to identify a real world **entity**, thus distinguishing one **entity** from one another.

Entity Definition - Data Values Heterogeneity (2)

Example: consider the Car entity represented in two different datasets A, and B.

Car in dataset A:

- Vehicle-ID: 1234
- Manufacturer: "Renault"
- Engine-type: "Electric engine"
- Fuel-type: "Electricity"

Car in dataset B:

- Vehicle-ID: ABCD
- Manufacturer: "Renault"
- Engine-type: "Electric engine"
- Fuel-type: "Electricity"

From the same source, we have two datasets in the same format, using the same structure of information. Nevertheless ..

- how can we know if the two car are the same entity or different ones ?
- is the identifier in dataset A equivalent to the identifier in dataset B ?
- the "Manufacturer" term in datasets A has the same meaning of "Manufacturer" in dataset B ?

Entity Definition - Activities

- It is necessary to handle the meaning heterogeneity to produce a KG(s) suitable to satisfy the initial purpose.
- To this end, the last phase of the iTelos methodology is structured in three different **activities**:
  - **Entity Matching**
  - **Entity Identification**
    - Identifiers
    - Identifying Sets
  - **Entity Mapping**

Entity Definition - Activities

- Entity Matching
- Entity Identification
  - Identifiers
  - Identifying Sets
- Entity Mapping

Entity Definition - Entity Matching

- The real world entities, **represented by their values**, can be represented through different properties, and properties values, within different **datasets**.
- This is known as **the entity matching problem**, and it has two main consequences:
  1. (Schema layer) The need to find **the right set of properties** between the different datasets where multiple representations of the same entity, can be present.
  2. (Data layer) The need to set **the correct property values**, if multiple representations share the same properties, but having different values.

Entity Definition - Entity Matching (2)

- It is important to notice that, if the previous phases have been performed by considering the iTelos **middle-out approach**, most of the misalignment between ETypes (teleontolgy) and Entities (datasets) should be solved.
  - This happens because the teleontology has been modeled by considering the datasets, and the datasets have been aligned with the modelling choices adopted in the teleontology.

Entity Definition - Entity Matching (3)

- Nevertheless, some of such misalignment could be present in this phase.
  - **For example**: an entity is present in two datasets A and B, but in dataset A the entity is well described by a rich set of properties, while, in dataset B, the entity appears described by one single property.
  - The entity matching problem needs to be solved by understanding if the two representation of the same entity correspond, and if the properties values can be matched.

Entity Definition - Entity Matching (4)

- **How to solve entity matching misalignment ?**
- A possible solution is provided by **Metadata**.
- In particular, thus metadata carrying information about the provenance and the reliability of the entities having conflicts.
  - **Author** and **Organization** metadata allow us to understand who created the data, thus giving us a criteria in order to decide which property/value should be considered, or not for the same entity.
  - **Creation Date** and **Modification Date**, similarly give us information about how much up-to-date the data are (too old or too new, depending by what our purpose requires).
  - Also for entity matching, **the purpose** (used to create the data we are reusing) is the main criteria to be used in order to solve conflicts.

Entity Definition - Activities

- Entity Matching
- Entity Identification
  - Identifiers
  - Identifying Sets
- Entity Mapping

Entity Definition - Entity Identification

- When the entity matching conflicts have been clarified, we need to **formally identify the different entities**.
- More in details, we need to:
  - identify an entity within a **single dataset**;
  - adopt the **same type of identification**, if the same entity is represented in two (or more) different ways, **within different datasets**.
- **How to formally identify the entities in the datasets ?**

Entity Definition - Entity Identification (2)

- An entity (like the etypes) is identified by its properties.
- Sometimes within (well formed, quality) datasets it is already present a specific property aiming at identifying the entity it belongs to.
  - Such a property is called **Identifier**.
- There are **multiple kinds of identifiers**, depending on how the entities need to be identified.

Entity Definition - Entity Identification - Identifiers

- **URI**: A Uniform Resource Identifier (URI) is a unique sequence of characters that identifies a logical or physical resource used by web technologies.
- A URI can be defined as:
  - **URL**: A Uniform Resource Locator (URL) is a URI that specifies the means of acting upon or obtaining the representation of a resource, i.e. specifying both its primary access mechanism and network location.
  - **URN**: A Uniform Resource Name (URN) is a URI that identifies a resource by name in a particular namespace.
  - Examples and more details can be found directly at Wikipedia URI
- Nevertheless, identifiers are not always provided in the datasets.

Entity Definition - Entity Identification - Identifying Sets

- When an identifier (a single entity's property) is not available, an entity can be identified uniquely by the union of the values from two (or more) of its properties.
  - Such a property composition is called **Identifying Set**.

Identifying Set: a set of etype's properties which, through their values, identify uniquely an entity (defined for such an etype) within the whole set of entity considered.

Entity Definition - Entity Identification - Identifying Sets

Bus in dataset A:

- Production-year: 2007
- Manufacturer: "Iveco"
- Model: "AX-123"
- Engine-type: "Electric engine"
- Fuel-type: "Electricity"

Bus in dataset B:

- Production-year: 2007
- Line-number: "13-A"
- Seats: 30
- Daily-travel-time: 650
- Model: "AX-123"

The Identifying Set (IS) is defined as follow:

$IS_{Bus} = Production\text{-}year, Model$

It allows the matching between the two *Bus* entities into a single one.

Entity Definition - Activities

- Entity Matching
- Entity Identification
  - Identifiers
  - Identifying Sets
- Entity Mapping

Entity Definition - Entity Mapping

- The last activity, called Entity Mapping, aims at concretely **merging** the information representation defined in the **teleontology**, with the relative information values in the **datasets**.
- The activity is composed by many **mapping operations** that concretely implement **the solution to the entity matching problem**.
- Moreover, a specific type of **mapping operation** is performed to concretely define **the identifiers for the entities**, to be considered in the final KG(s).
- The Entity Mapping activity is performed by using the **Karma** tool.

Entity Definition - Mapping Operations

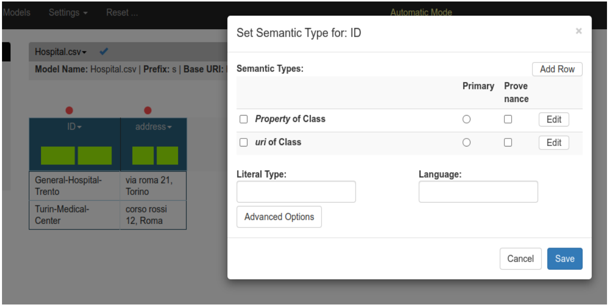
- An example of mapping operation using Karma.

Patient .csv UTF-8

Model Name: Patient .csv | **Prefix:** s | **Base URI:** http://localhost:8080/source/ | **Github URL:** disabled

ID	Name	date-of-birth	address	Gender	Doctor
2	Bruce Banner	19901203	via roma 2, Trento	Male	Doctor-1
1	Tony Stark	19860511	via verdi 12, Roma	Male	Doctor-2
4	Anna Verdi	19750903	via castello 2, Tornio	Female	Doctor-1
3	Lucia Skywalker	19820713	piazza venezia 1, Palermo	Female	Doctor-2

Entity Definition - Mapping Operations - URI definition

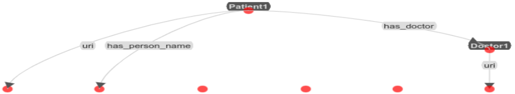


Entity Definition - Mapping Operations - URI definition

ID	Name	Surname	Hospital
Doctor-1	Mario	Draghi	General-Hospital-Trento
Doctor-2	Clara	Bella	Turin-Medical-Center

Patient .csv

Model Name: Patient .csv | **Prefix:** s | **Base URI:** http://localhost:8080/source/ | **Github URL:** disabled



ID	Name	date-of-birth	address	Gender	Doctor
2	Bruce Banner	19901203	via roma 2, Trento	Male	Doctor-1
1	Tony Stark	19860511	via verdi 12, Roma	Male	Doctor-2
4	Anna Verdi	19750903	via castello 2, Tornio	Female	Doctor-1
3	Lucia Skywalker	19820713	piazza venezia 1, Palermo	Female	Doctor-2

Entity Definition - The final Knowledge Graph

- The output of the Entity Mapping activity twofold:
  - **The mapping model**: a RDF-Turtle (ttl) file defining all the mapping operations performed using the Karma tool.
  - **The KG(s)**: one, or a set of RDF-Turtle (ttl) files defining the main output of the last iTelos phase.

Phase 5 - Entity Definition - Summary

- In the last iTelos phase we do:
  - the handling of the data values heterogeneity, by:
    - Entity matching and,
    - Entity identification.
  - The merging of the knowledge and data layer, handled during the previous iTelos phases.
  - The generation of the final process output.