KDI : Knowledge and Data Integration

Semantic Data Alignment

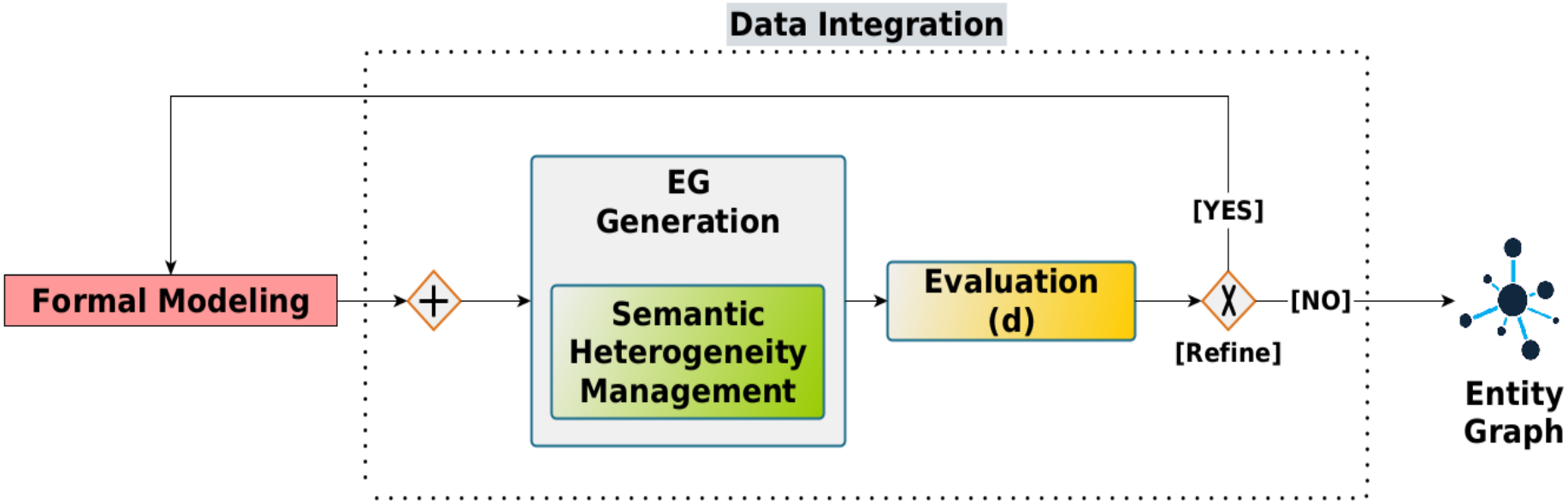
iTelos Data Integration Phase

Simone Bocca

Contents

Contents

Data Integration phase



Data Integration objective

Data Integration is the fourth iTelos phase

Inputs:

- ETG.
- Dataset syntactically aligned.

Outputs:

- EG.
- Metadata.
- Project documentation.

The Data Integration phase aims to build the final EG, populating the ETG previously produced with the datasets entities. In this phase the knowledge and data layer are, in the end, merged together and the data semantic heterogeneity is handled.

Contents

Semantic Heterogeneity

The Data Integration phase has to handle with the semantic heterogeneity within the datasets collected.

Q: What is the data semantic heterogeneity ?

A: Such kind of heterogeneity indicates the presence of multiple representation of the same real world entity.

The semantic heterogeneity is a:
"consequence of the more general phenomenon of the diversity of the world and of the world descriptions." (Giunchiglia, Fumagalli 2020)

Semantic Heterogeneity - Example

Consider two different dataset, A and B, both including entities representing buses.

Bus in dataset A:

- Vehicle-ID: 4321
- Manufacturer: "Iveco"
- Engine-type: "Electric engine"
- Fuel-type: "Electricity"

Bus in dataset B:

- Vehicle-ID: 4321
- Line-number: "13-A"
- Seats: 30
- Daily-travel-time: 650

In dataset A, a bus is seen from the point of view of the manufacture, while in dataset B the same bus is seen from the point of view of the public transportation company.

The same real word entity is represented using different properties due to the **different function associated to the entity** within the two datasets (i.e., in dataset A the bus is a motor vehicle, while in dataset B is a transportation vehicle).

Contents

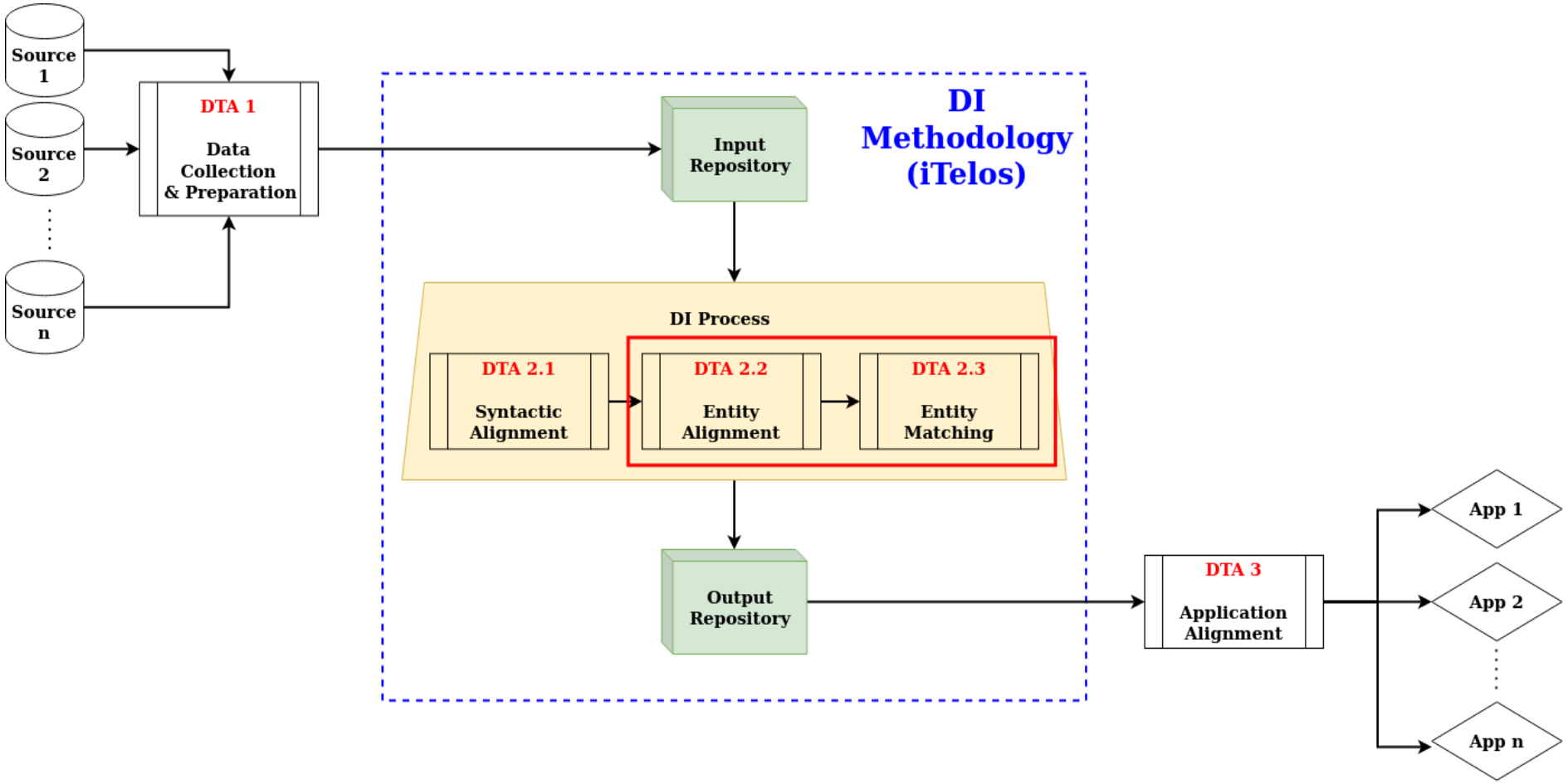
Address Semantic Heterogeneity

Q: How to handle with such kind of heterogeneity within the datasets ?

A: iTelos, with the objective to cover the semantic misalignment, defines two separate activities:

- **Entity alignment (DTA - 2.2)**: which aims to map multiple entity representation (between different datasets) to the single, purpose-specific schema (ETG).
- **Entity matching (DTA - 2.3)**: which aims to identify if different entities in the datasets can represent the same real world entity, and as a consequence, should be merged together within the final EG.

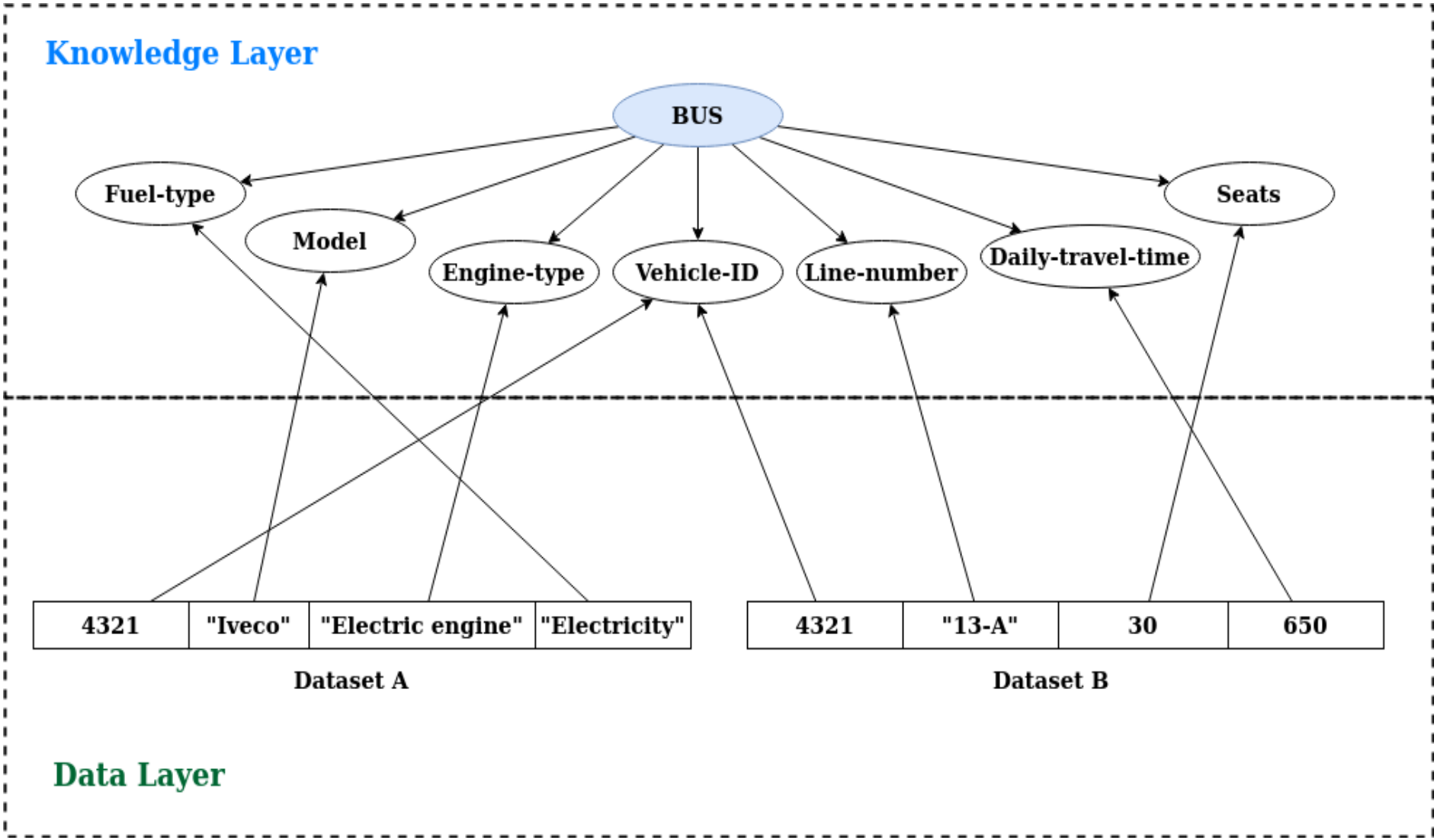
iTelos DTA - 2.2, 2.3



Entity alignment

- The entity alignment activity is responsible for the merge of the two iTelos layers, knowledge and data.
- In this activity the Data Scientist (DS), using a specific tool, maps the datasets values (representing entities and entity properties) on the etypes and etype properties defined in the ETG.
- Knowing the data meaning, the DS can map the datasets values on the rights ETG's etypes and properties, covering in this way the semantic misalignment between the datasets.

Mapping operations - Example



Mapping operations - Notes

- **Entity identification**: The datasets can include values used to identify the single entities, like *4321* in the previous example. Such values are used as *identifiers* to distinguish between the entities across different datasets.
- But the identifiers are not always present (and/or specified) within the datasets ...
- The lack of identifiers, can lead to add in the EG, separate entities which instead refer to the same real entity.
- This is why the **Entity Matching activity** is required.

Entity matching

The Entity Alignment activity produce a first version of the EG, where the issues mentioned in the previous slide, have yet to be addressed.

Q: How to match two separate entities that should be a single one ?

A: A solution to identify the entities, can be considered at knowledge level, by defining, for each etype in the ETG, an **Identifying Set**.

> *Identifying Set*: a set of etype's properties which, through the values associated to them, identify uniquely an entity (defined for such etype) within the whole set of entity considered by the DI process.

Entity matching - Examples

Consider the same two datasets as before where the identifiers are not available anymore.

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 defined as follow:

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

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

Entity matching - Merge conflicts

Once the entity matching has been identified, the attributes of the entities to be merged, are checked in order to discover possible conflicts while merging the two entities.

For example, considering the two Bus entities as follows:

Bus in dataset A:

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

Bus in dataset B:

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

Q: Which value of the property *Color* should appear after merging the entities ?

Entity matching - Merge conflicts (2)

Such kind of conflicts between entities that have to be merged, can be caused by errors within the dataset or different representation of the same entity (due to different context in which the datasets have been created).

> **A**: Specific kind of metadata associated to the datasets and/or directly to the entities within the datasets, can help to decides a solution to solve conflicts like the one presented in the previous slide. One of them is the **Provenance**.

The Provenance is a metadata indicating the origin of the dataset/entity/value. Thanks to this information the data scientist can select which is the most appropriate values to associate to the etype property that produce the conflict.

Moreover the Provenance improve the level of reliability of the resources considered.

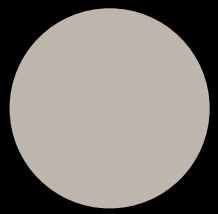
Contents

Summary

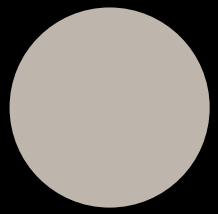
In this lecture we discussed:

- The main objectives of the Data Integration phase.
- The semantic heterogeneity and how to deal with it.
- Entity alignment and matching activities.
- The Provenance support to the entity merging procedure.

KDI : Knowledge and Data Integration



Simone Bocca



Semantic Data Alignment

iTelos Data Integration Phase