KDI : **Knowledge and Data Integration**

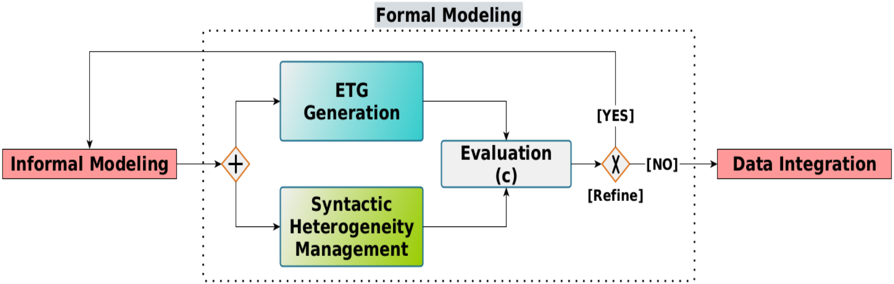
Syntactic Heterogeneity Management

iTelos Formal Modeling Phase

Simone Bocca

Contents

Formal Modeling phase



Contents

1. **Syntactic Heterogeneity**

2. Data misalignments

3. Data and Knowledge layers alignment

4. Summary

Syntactic Heterogeneity

The second main objective of the Formal Modeling phase is to handle the syntactic heterogeneity within the datasets collected.

Q: What is the data syntactic heterogeneity ?

A: Such kind of heterogeneity appears at data value level, when data use different values to represent the same kind of information.

Syntactic Heterogeneity

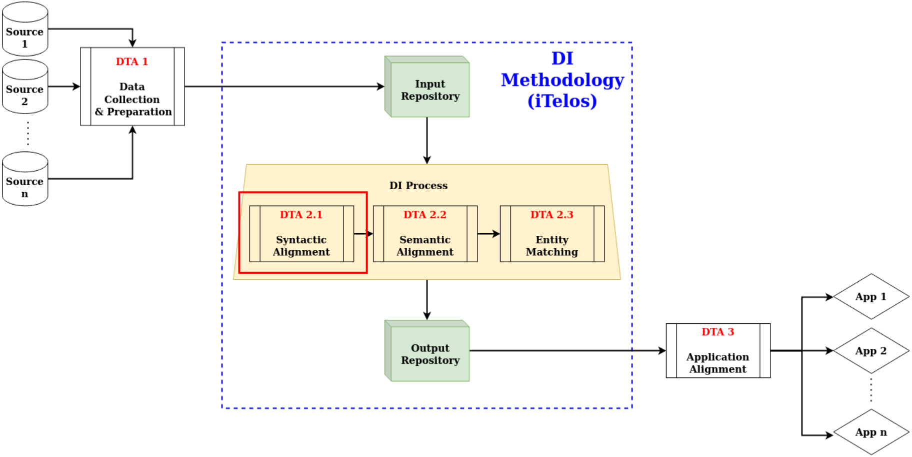
Q: How many types of data value misalignment we can find within the datasets ?

A: There are in general three types of misalignment to consider addressing the syntactic data heterogeneity:

- Data types misalignment
- Data value format misalignment
- Data value language misalignment

The **Data Transformation Activity - 2.1** aims to deal with the above listed misalignments, addressing in this way the syntactic heterogeneity.

iTelos DTA - 2.1



Contents

1 Syntactic Heterogeneity

2 **Data misalignments**

3 Data and Knowledge layers alignment

4 Summary

Data types misalignment

The data types misalignment appears when the same information is represented using different data types.

Example: consider two different dataset, A and B, containing data about students in Trento.

Entity in dataset A:

- Name: "Simone"
- Surname: "Bocca"
- Age: 29
- Telephone: 3328877451

Entity in dataset B:

- Name: "Simone"
- Surname: "Bocca"
- Age: "29"
- Telephone: "3328877451"

Data types misalignment

The data types misalignment appears when the same information is represented using different data types.

Example: consider two different dataset, A and B, containing data about students in Trento.

Entity in dataset A:

- Name: "Simone"
- Surname: "Bocca"
- Age: 29
- Telephone: 3328877451

Entity in dataset B:

- Name: "Simone"
- Surname: "Bocca"
- Age: "29"
- Telephone: "3328877451"

Age and *Telephone* values, in dataset A are represented as Integer and Long values. While in dataset B the same information is represented using Strings.

Data value format misalignment

The data value format misalignment appears when different formats of the same data type are adopted for same information in different datasets.

Example: consider the same two dataset, A and B.

Entity in dataset A:

- Name: "Simone"
- Surname: "Bocca"
- Age: 29
- BirthDate: "1992-08-13 15:35:03"

Entity in dataset B:

- Name: "Simone"
- Surname: "Bocca"
- Age: 29
- BirthDate: "713720103"

Data value format misalignment

The data value format misalignment appears when different formats of the same data type are adopted for same information in different datasets.

Example: consider the same two dataset, A and B.

Entity in dataset A:

- Name: "Simone"
- Surname: "Bocca"
- Age: 29
- BirthDate: "1992-08-13 15:35:03"

Entity in dataset B:

- Name: "Simone"
- Surname: "Bocca"
- Age: 29
- BirthDate: "713720103"

BirthDate value, in dataset A is represented as a String in ISO date format. While in dataset B the same information is represented using a unix timestamp String.

Data value language misalignment

The data value language misalignment appears when different natural languages are adopted for same information in different datasets.

Example: consider the same two dataset, A and B.

Entity in dataset A:

- Name: "Simone"
- Surname: "Bocca"
- Age: 29
- Student-type: "Doctoral student"

Entity in dataset B:

- Name: "Simone"
- Surname: "Bocca"
- Age: 29
- Student-type: "Studente di dottorato"

Data value language misalignment

The data value language misalignment appears when different natural languages are adopted for same information in different datasets.

Example: consider the same two dataset, A and B.

Entity in dataset A:

- Name: "Simone"
- Surname: "Bocca"
- Age: 29
- Student-type: "Doctoral student"

Entity in dataset B:

- Name: "Simone"
- Surname: "Bocca"
- Age: 29
- Student-type: "Studente di dottorato"

Student-type value, in dataset A is represented as a String in English. While in dataset B the same information is represented using an Italian String.

Contents

1 Syntactic Heterogeneity

2 Data misalignments

3 **Data and Knowledge layers alignment**

4 Summary

Data and Knowledge layers alignment

Also in the Formal Modelling phase, iTelos maintains the synchronisation between knowledge and data layer.

All the syntactic alignments are performed considering how the information is modelled in the ETG. In other words **the knowledge layer leads the actions to fix the data misalignment**.

Contents

1 Syntactic Heterogeneity

2 Data misalignments

3 Data and Knowledge layers alignment

4 **Summary**

Summary

In this lecture we discussed:

- What is the syntactic heterogeneity and when it appears in data resources.
- Which are the possible types of data misalignment to consider addressing the syntactic heterogeneity.
- How the methodology support the syntactic data alignment with the help of the knowledge layer.

KDI Knowledge and Data Integration

Simone Bocca

Syntactic Heterogeneity Management

iTelos Formal Modeling Phase