KDI : **Knowledge and Data Integration**

Data Collection & Preparation

iTelos Inception Phase

Simone Bocca

Contents

1. Data Collection

2. Data Preparation

3. Summary

Contents

1 **Data Collection**

2 Data Preparation

3 Summary

Resource collection scope

The second main objective of the Inception phase is to collect the data, and knowledge, resources that have to be integrated to produce the final EG.

A data integration process, in general, considers resources (both teleologies and datasets) which can bring the following two results:

- Increase the number of **entities/entity types**
- Increase the number of **entity attributes/entity type properties**

Q: How to choose the right resources ?

A: The user's Purpose, transformed in a list of CQs, leads the choice of which resources have to be collected.

Heterogeneous Data sources

While the knowledge resources are usually defined using few standards formats and languages (such as RDF and OWL) the data resources come from different data sources and have been produced for different purposes, due to that they present an high level of **heterogeneity**.

In the Inception phase the required datasets have to be collected and such heterogeneity appears in terms of **different data formats** adopted by the selected data sources.

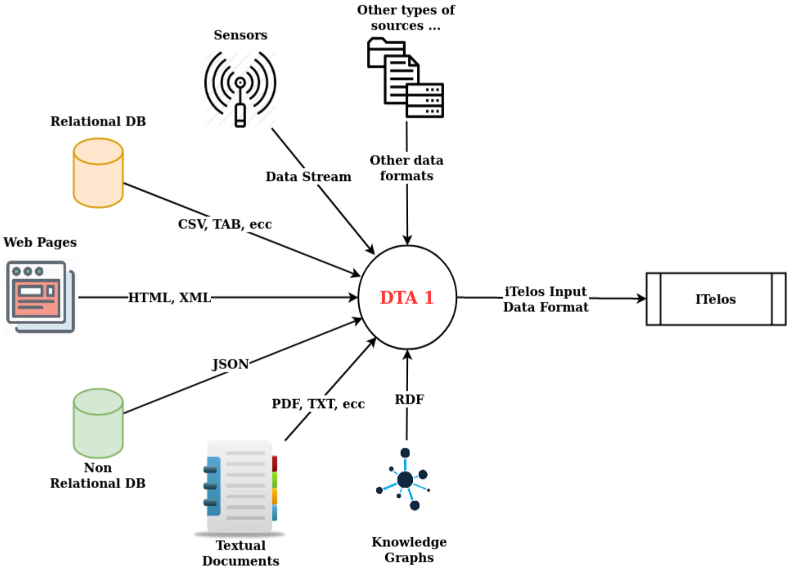
Data sources types

We can divide the possible data sources in the following categories, based on the kind of data they provide:

- **Structured**: CSV, TAB, JSON, Spreadsheet, XML, and others.
- **Semi-structured**: web pages (HTML), data obtained by web scraping.
- **Unstructured**: textual document (PDF, txt).
- **Media**: images, videos.
- **Streams**: continuous data obtained from sensors.

Q: How to address this kind of heterogeneity ?

Address input heterogeneity



Contents

1 Data Collection

2 Data Preparation

3 Summary

iTelos Input Data Format

A: iTelos define its own input data format in order to consider, during the DI process, resources having the less possible input heterogeneity.

The iTelos input format allows (for the current methodology version) resources defined using one of the following structured data formats:

- Knowledge resources:
  - RDF-OWL
- Data resources:
  - CSV
  - Excel Spreadsheet
  - JSON
  - XML

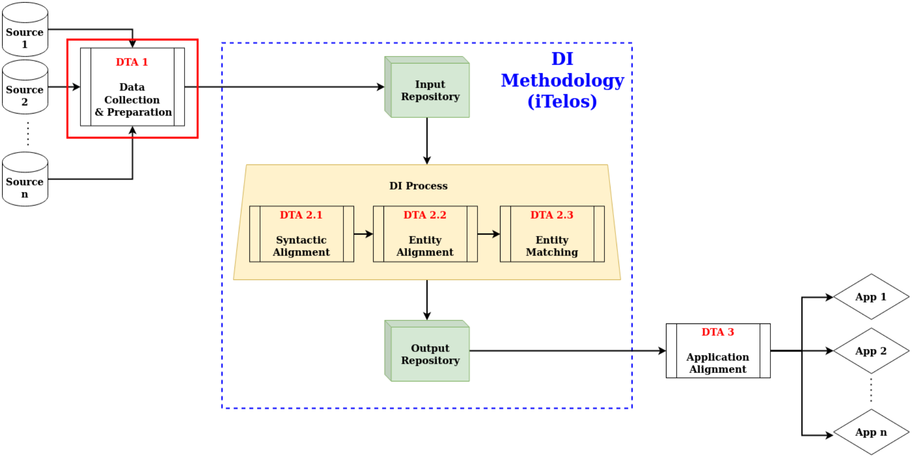
Data Transformation Activity 1
Input resources

The DTA-1 is the activity which receives in input the heterogeneous input resources, and aims to align such resources with the standards defined by the *iTelos input format*.

The DTA-1 considers also the possibility to deal with sensible data. For this reason the DTA-1 internal process is divided in two steps:

1. **Data normalization**: The first step aims to transform the datasets in order to be compliant with the iTelos input format. This step involves the usage of data conversion tools, libraries and/or dedicate scripts.
2. **Data anonymization**: The second step is required only if sensible information (like personal data) are included in the datasets collected. The anonymization aims to allow the DI process to manage the resources without privacy issues, and also to produce resources that can be shared with different projects.

iTelos DTA - 1


Source 1
Source 2
Source n
DTA 1
Data Collection & Preparation
Input Repository
DI Methodology (iTelos)
DI Process
DTA 2.1
Syntactic Alignment
DTA 2.2
Entity Alignment
DTA 2.3
Entity Matching
Output Repository
DTA 3
Application Alignment
App 1
App 2
App n

DTA-1 and Reusability

Observation

The resources transformed by DTA-1 not only are suitable to be used within the specific project for which they were collected, but also become exploitable for any other DI project which follows the iTelos methodology, in other words for different purposes.

This means that, in line with the basic iTelos principles, the DTA-1 improves the **reusability** of the resources considered by the DI project.

Contents

1 Data Collection

2 Data Preparation

3 **Summary**

Summary

In this lecture we discussed:

- The resource collection general criteria to be considered in Inception phase.
- The input heterogeneity derived from the different kinds of data sources.
- How to address the input heterogeneity.
- The first Data Transformation Activity.

KDI Knowledge and Data Integration

Simone Bocca

Data Collection & Preparation

iTelos Inception Phase