KDI : Knowledge and Data Integration

iTelos Principles

Data Integration Methodology Basic Principles

Simone Bocca

Contents

Contents

Data Integration (DI) Context

Q : Why, and where, we need a DI methodology ?

A : A DI methodology is a set of activities allowing its users to handle heterogeneous data, from different sources, in order to let them suitable to be exploited by one or more specific applications.

DI Methodology (iTelos) Life Cycle (1)

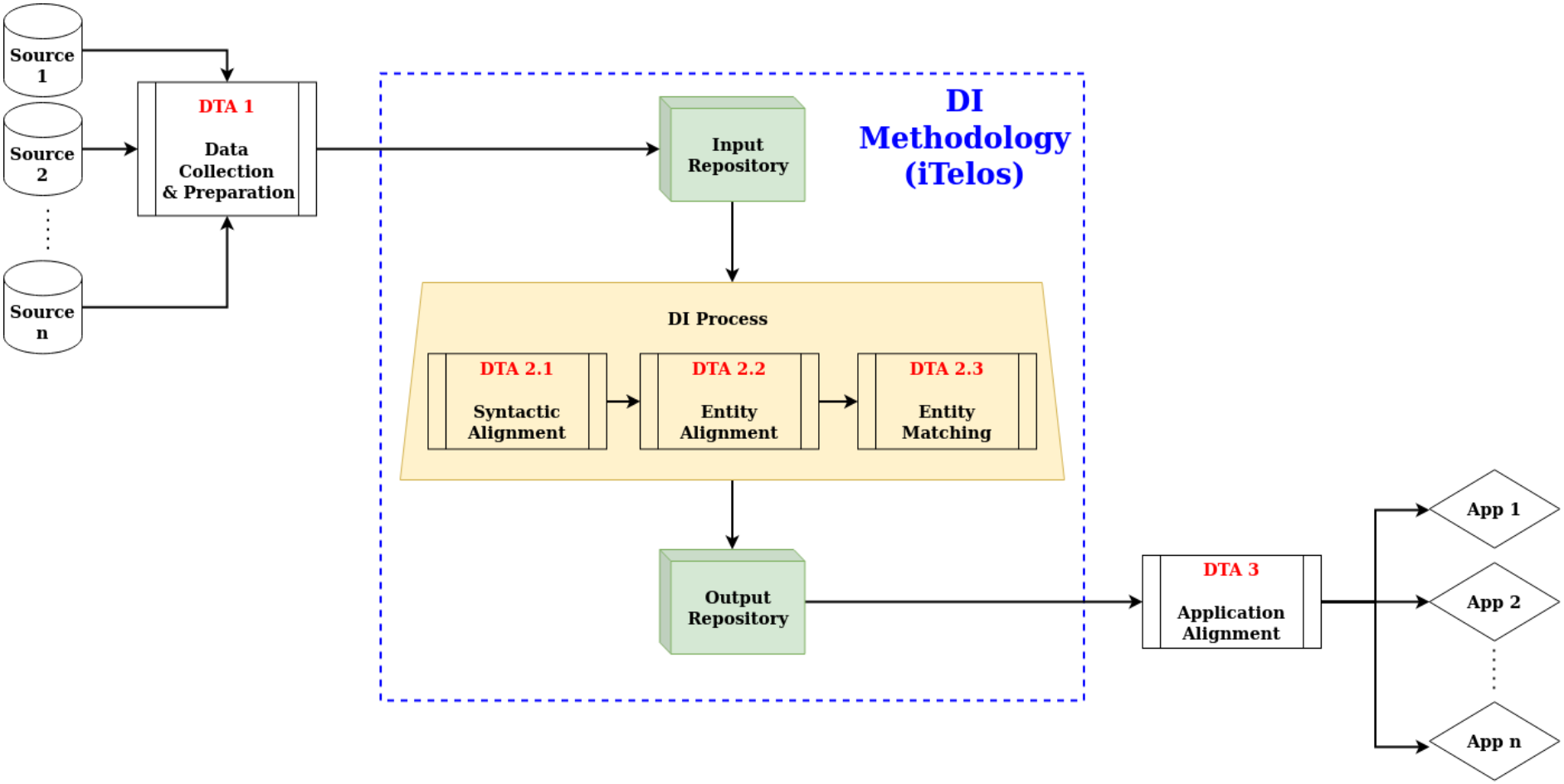
The DI methodology has to deal with multiple kinds of data sources:

- tabular datasets;
- object datasets;
- data streams;
- and others ...

and, it has to provide integrated data suitable to be exploited by different kinds of applications which adopt their own data formats, and standards, to deal with data. Moreover, in order to execute properly each activity involved in the methodology, the data have to be aligned using a single specific standard adopted internally by the methodology itself.

For these reasons, within the set of DI activities, we have to consider 4 different *Data Transformation Activities* (DTAs), plus and additional application specific DTA.

DI Methodology (iTelos) Life Cycle (2)



DI Methodology (iTelos) Life Cycle (3)

- **Data Collection & Preparation** (DTA-1): align the different source's data formats, and data standards, representing information through a single data standard. The aligned data can be then collected within the methodology's *Input Repository*.
- **Syntactic Alignment** (DTA-2.1): align the data value formats by adopting the same data standards for similar data types.
- **Entity Alignment** (DTA-2.2): align the semantic of the data (*entity schema* representation and *word sense disambiguation*).
- **Entity Matching** (DTA-2.3): align modeled entities with already existing representations of the same entities.
- **Application Alignment** (DTA-3): the last transformation, considered as out of scope for the DI methodology, aims to align the integrated data in order to let them suitable to be used by a specific application.

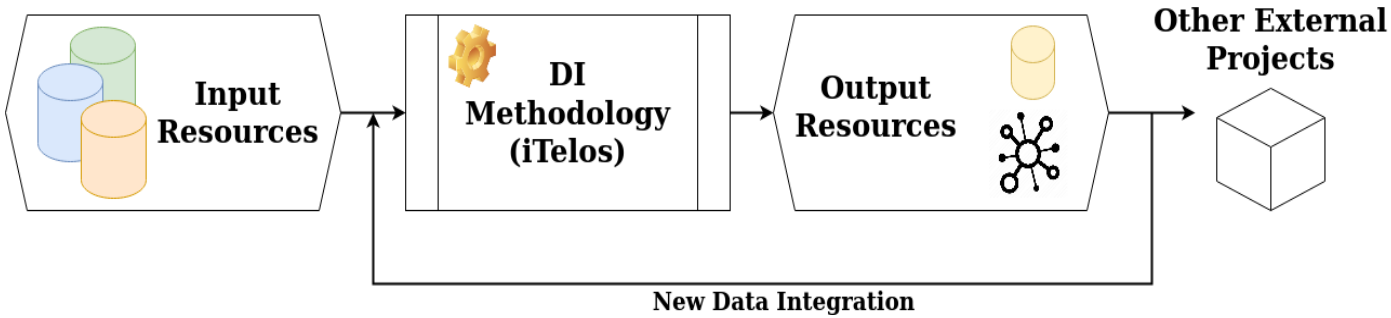
Contents

Reusability & Shareability

"It is a fact that, when developing a new application, it is virtually impossible to reuse, as-is, existing datasets. This difficulty is the cause of additional costs, with the further drawback that the resulting application will again be hardly reusable." (Giunchiglia, 2021)

Q: How to break the loop described above ?

A: Enhancing the reusability and shareability of the resources involved and produced using the DI methodology.



Reusability - Why

- Reuse resources, instead of produce (and/or pay for) new ones, reduce the integration effort in terms of:
  - computational overhead of the integration process;
  - time consumed to perform all the integration activities;
  - cost of resources. Often, dealing with huge quantity of data and/or with high quality data, the cost of resources can have a strong impact on the feasibilty of application which has to exploit DI outcomes.

- The level of heterogeneity decrease when the resources are reused instead of create new instance of them.

Reusability - How

- Selection of resources with high shareablity (see next slide).
- DTAs designed considering multiple data formats and standards;
- Resource classification, in order to identify the most (and less) reusable resources. This improves the identification of already existing resources which can be used during the integration process.
- Use of *Metadata* to enrich the information carried by the data allowing, in this way, the methodology users to better identify the most reusable resources.
- Use of dedicated evaluation activities, included in the DI process, to ensure a proper level of reusabiltiy of already existing resources.

Shareability - Why

- In order to reuse the resources produced by the DI methodology, they have to be created enhancing their sharability. In other words the capability to be shared among different projects with different purposes.

- The shareability of resources produced support the *Composition of Domains* (see Slides XX), namely, the usage of resources produced in one or more specific domains of interest, to represent a larger domain, by composition.

Shareability - How

- KGs designed considering as much as possible already existing, well formed, knowledge resources.
- Knowledge resources produced using the Teleology theory (see Slides Foundational Teleology).
- Dedicated evaluation activities included in the DI process to ensure a proper level of shareability in the resources produced.

Contents

Teleologies

> Teleology is the study of ends and goals, things whose existence or occurrence is purposive.

Concretely, in our context, teleologies are ontologies but with the proviso that teleologies focus on function and on how a chosen representation fits a certain purpose.

In other words, the teleologies are the way adopted in the DI methodology (iTelos) to model (design, represent) the information that needs to be exploited by final users.

Teleologies - Why

The usage of teleologies align the representation of information coming from different sources, using the same foundational theory to model the causality of the world. This approach allows:

- reduction of semantic heterogeneity;
- production of high shareable knowledge resources;
- compositionality of domains modeled through teleology (See Domain composition principle)

Contents

Composition of Domains

Domain of Interest (DoI): The portion of the world that involves all the information elements used to satisfy a specific purpose.

Using KGs to integrate data implies the capability to represent, within the KG itself, the DoI relative to the data integration purpose.

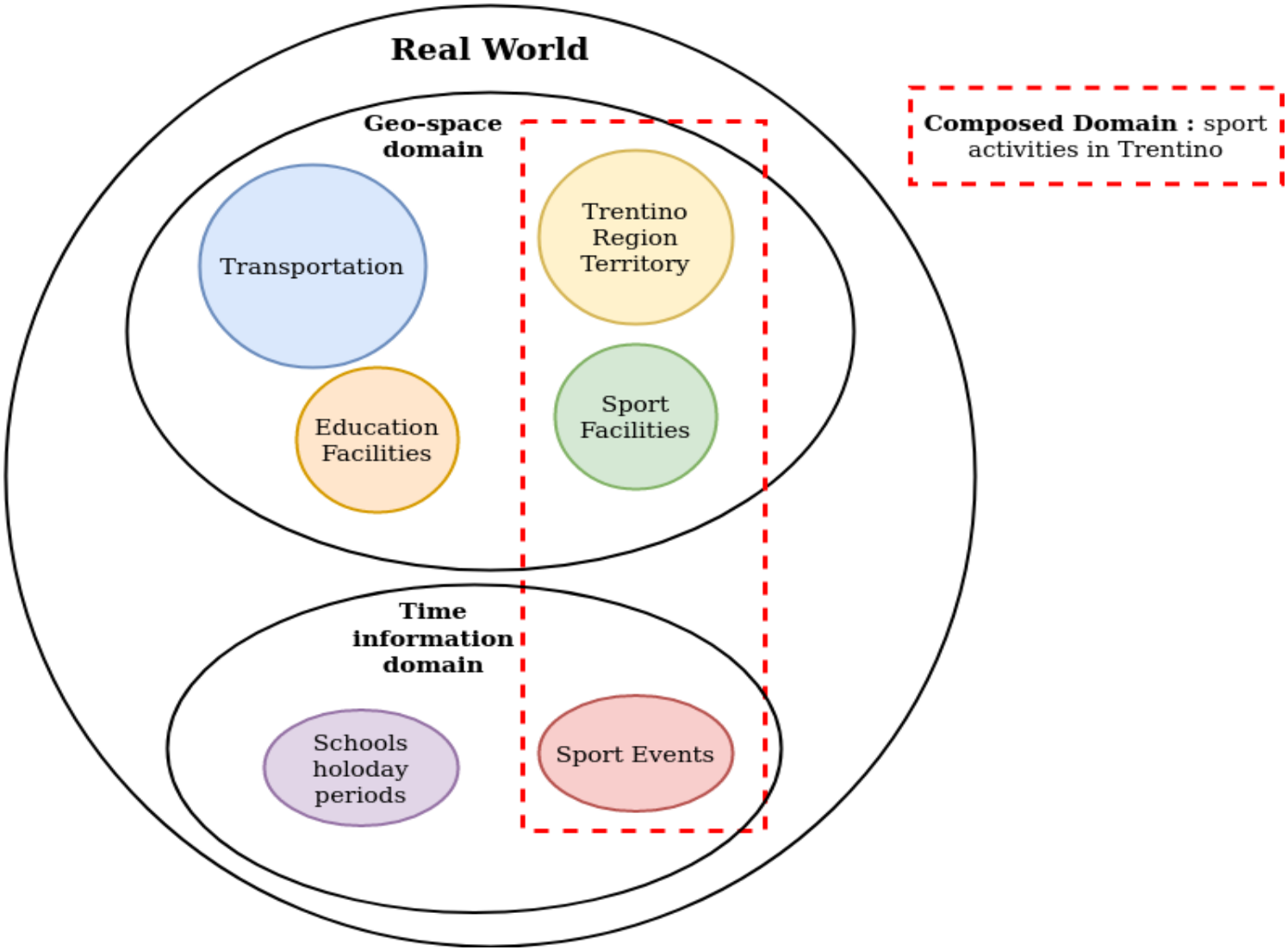
HOW ? Using knowledge resources (Teleologies)

> The iTelos DI methodology, exploiting the Teleology theory, allows the representation of purpose specific DoI by composition of domains designed using high shareable resources.

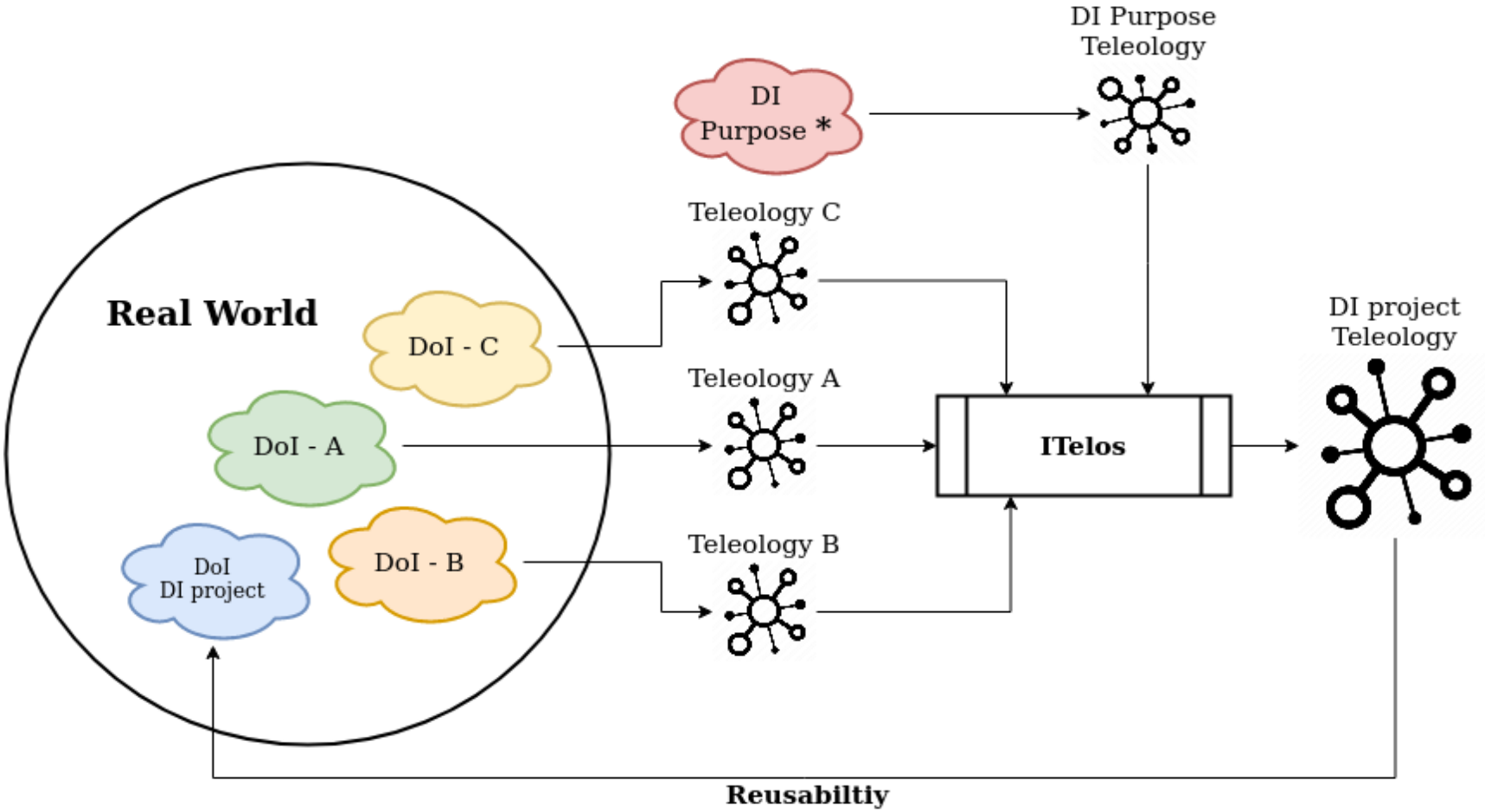
ADVANTAGES :

- Reuse of existing knowledge resources.
- Reduced effort in common aspect design
- Improve the number of possible DoI which can be composed.
- Interoperability among projects/applications.

Composition of Domains - Example



Composition general process



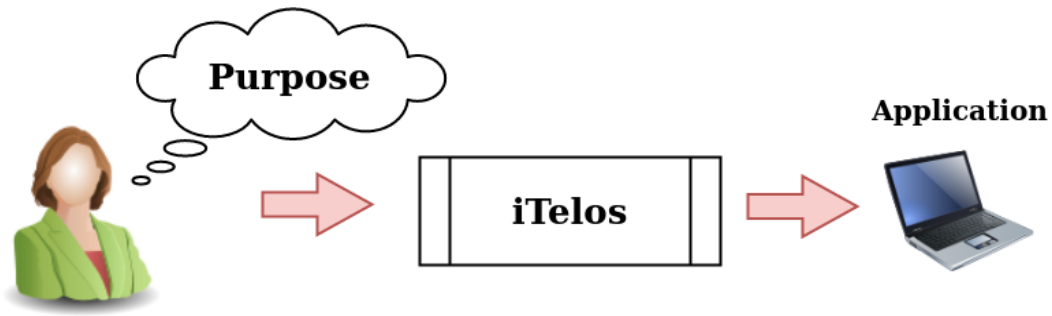
* See next slides for DI Purpose definition.

Contents

Data Integration Purpose

DI Purpose: The objective to be achieved exploiting the DI result.

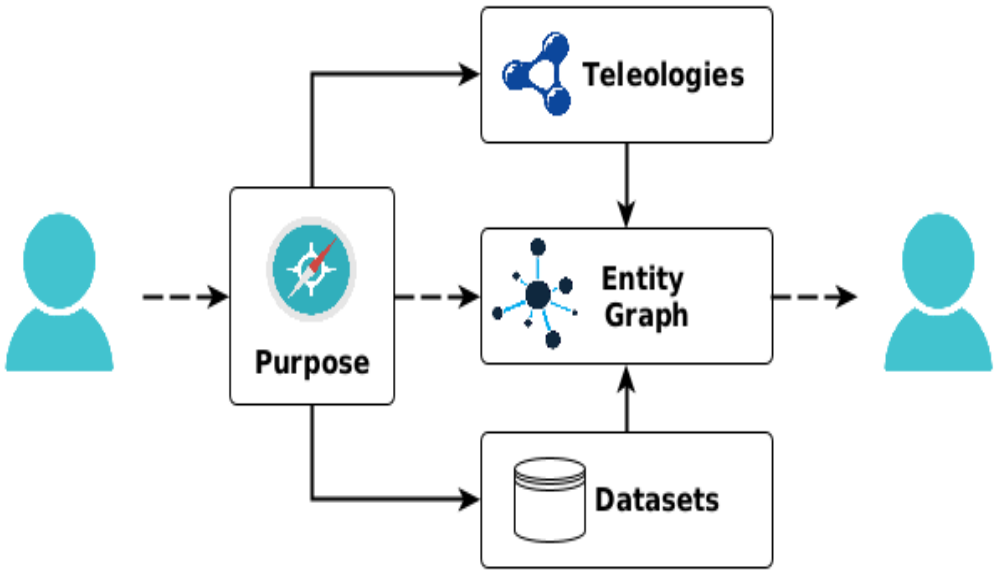
The DI Purpose, expressed by the final user as a natural language sentence, define the main goal for the whole data integration. It represent what the final user should be able to do exploiting the DI final outcome (KG). Due to that, the Purpose leads the whole integration process and makes iTelos be a **Purpose driven DI methodology**.



Purpose driven DI methodology

The Purpose leads the mains methodology aspects:

- Project domain of interest definition.
- Resources collection (both data and knowledge resources).
- Datasets transformation.
- Project teleology design.



Contents

Knowledge & Data Alignment

A KG, as result of a DI process, is composed by knowledge (teleologies) and data (datasets) resources combined together in the best way possible in order to achieve the user's Purpose. There are three different approaches to build such kind of KGs:

1. **Knowledge centric**: The data schema design comes first and then the data ar e aligned to it.
2. **Data centric**: The data schema is mainly (some adaptation are always required) extracted from the data to be integrated.
3. **Middle-out approach**: The data schema (teleology) is designed considering already existing knowledge resources **AND** the data. While the datasets are adapted to the schema designed.

Knowledge & Data Alignment - Limitations

These first two approaches brings some limitations:

1. The knowledge centric approach tends to produce KGs too general, being focused on the knowledge layer and less on the data which carry the information to be exploited.

2. The data centric approach tends to produce KGs too data specific, building the schema only from the datasets. Moreover, in this approach, the reuse of knowledge resources is less considered.

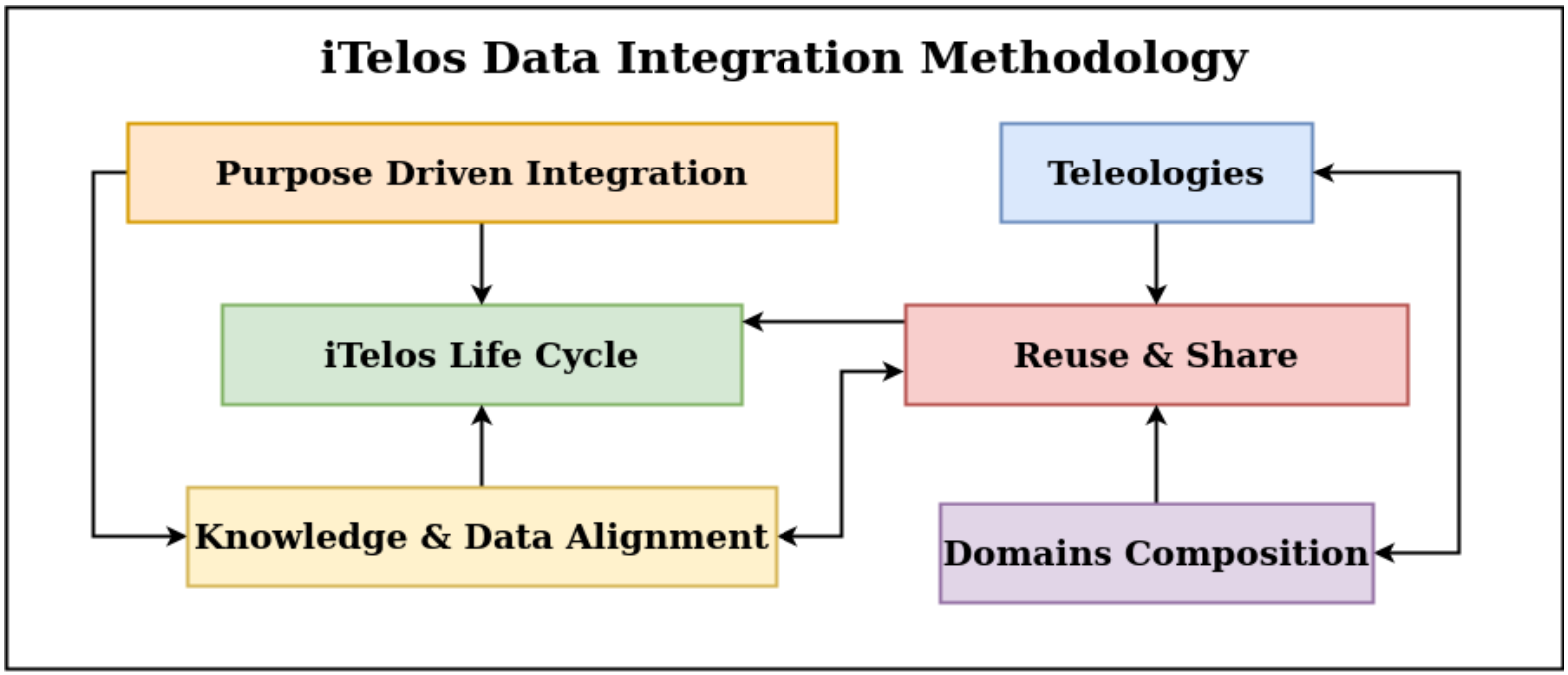
Knowledge & Data Alignment - Middle-out approach

The third approach instead, adopted by iTelos, defines a trade off providing a possible solution for the limitations listed above.

During the whole iTelos DI process the knowledge and data resources are always **considered in parallel**, aligning the management of the former with the management of the latter, in order to adapt the schema to the data and vice versa. To help this approach the methodology includes different **evaluation activities** which aim to test a correct alignment between knowledge and data layer.

Contents

iTelos Principles - Summary



A → B "A support B"

KDI : Knowledge and Data Integration

Simone Bocca

iTelos Principles

Data Integration Methodology Basic Principles