Part 5.2
iTelos

1 KG Construction

2 iTelos

3 Phase 1 - Purpose Definition

4 Phase 2 - Information Gathering

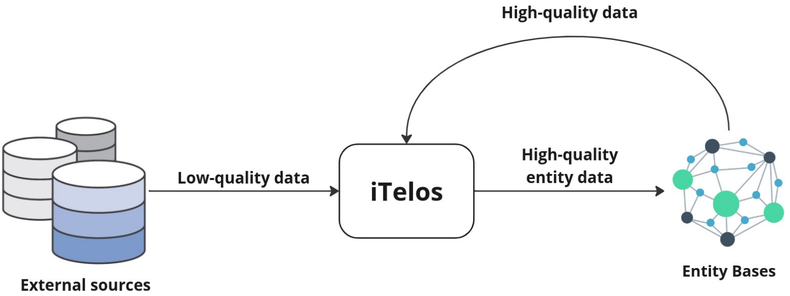
5 Phase 3 - Language Definition

6 Phase 4 - Schema Definition

7 Phase 5 - Entity Definition

A KGC Methodology based on Data Reuse

- We aim to cover the lack of a methodology for KGC with **iTelos** [31].
- iTelos is a methodology for KGC, which aims at **reducing the cost** of the entire process **by maximizing the reuse** of already existing resources.
- The key idea: **circular data economy**

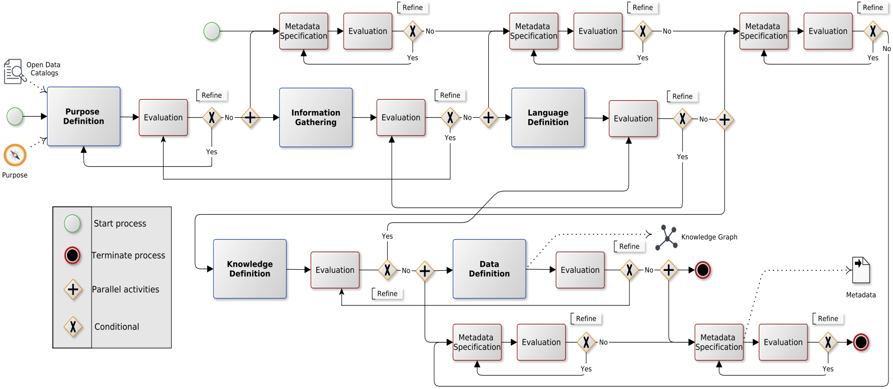


[31] "I" to indicate **"my"**, plus "Telos" that from the greek means **"purpose"**

iTelos - Circular Data Economy

- The iTelos methodology approach consists in two phases:
  - **First Iteration**: iTelos takes in input low quality data, and transforms such data into high quality entity KGs.
  - **Next iterations**: iTelos takes in input the low quality data (as before) plus the KG produced by the previous iTelos executions.
- The processing cost (reuse cost) of the KGs produced by iTelos **is less than the cost of reuse low-quality data**.
- By scaling the iTelos approach in future, the reuse of high quality KGs **will replace** the reuse of low-quality data, thus **reducing the cost of the overall KGC process**.

iTelos Methodology - The Full Structure



Purpose Definition

- The first iTelos phase is focused on the concrete definition of **”The Purpose”**.

- The purpose is **the objective of the user** that is using the methodology.
  - The idea, the why the user is constructing a (or a set of) KG(s).

- It can be seen as **the set of requirements** to be formalized, that will be used to test and validate the iTelos outputs (final and intermediate for each phase).

- However, the purpose implicitly include the ”user point of view”.
  - **By defining the purpose, the user will also represent her way to represent the real world entities!**

Purpose Definition

- **Input**: a natural language sentence representing the user's Purpose (plus, optionally, a list of already identified data sources to be exploited).
- **Output**: a set of documents and models in which the Purpose's requirements are extracted and formalized.
- **Objective**: to formalize the functional requirements implicitly included in the input user's purpose.

Information Gathering

- **Input**: a set of data sources identified previously, plus the formalized user's purpose.
- **Output**: a set of resources collected from the input data sources, suitable to satisfy the purpose.
- **Objective**: the second phase of iTelos aims at collecting the resources, to be processed, with the objective to build the final KG(s)

Information Gathering

- The gathering of information includes the collection of resources **for all the diversity layers**: data, knowledge and language.
- Notice how, depending by the source from which the resources are collected, they have different levels of quality:
  - **Low-quality sources**: the resources are collected from the existing sources, which not necessarily adopt iTleos, thus the quality level is, in average, lower.
  - **iTelos data catalogs**: the resources collected have been produced by iTelos executions, thus the quality level is, in average, higher.

Language Definition

- In this phase, iTelos aims at defining the "**language of the KG(s)**".
  - concepts and terms used to define the information to be exploited
- Notice that the information in the KG(s) could be defined by using not only **natural languages** but also **domain languages**.
  - standard concepts and terms defined for a specific domain (e.i, healthcare standards, unit of measure codes).
- The language definition phase is supported by the UKC project [32]

[32] http://ukc.disi.unitn.it/

Language Definition

- **Input**: the resources collected previously, plus the formalized user's purpose.
- **Output**: a set of language resources defining the concepts and terms to be adopted to define the KG(s) information.
- **Objective**: the third phase of iTelos aims at fixing the right concepts and terms for the KG(s)'s information, thus reducing the semantic heterogeneity of the final outcome.

Knowledge Definition

- Once the information is clearly defined by fixed concepts and terms, it needs to be **structured**.
- The modeling of the knowledge layer of the KG(s) **unifies the representation** of the information handled by the KG(s)
- iTelos models the KG(s)'s structure by exploiting a precise knowledge modeling methodology (KTelos) (detailed in the next chapter) based on the teleontology and teleology theory.

Knowledge Definition

- **Input**: the resources previously collected (knowledge and data) and produced (language), plus the formalized user's purpose.
- **Output**: one, or a set of (one for each KGs to be produced) knowledge resources.
- **Objective**: the knowledge resources produced in this phase aims at:
  - unifying the representation of the information;
  - improving the **interoperability** and **reusability** of the final KG(s), by building knowledge resources reusing as mush as possible well-known standard domain ontologies and data schema.

Entity Definition

- **Input**: the resources previously collected and produced (knowledge, language and data), plus the formalized user's purpose.
- **Output**: the final KG(s).
- **Objective**: the last phase of the methodology aims at merging the knowledge resources previously defined, with the cleaned and formatted data to be considered by the KG(s), thus producing the final concrete outcome.

Entity Definition

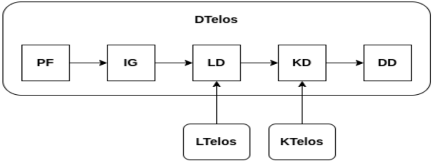
- The last phase of the methodology is dedicated to the data layer of the final KG(s).
- It is supported by a specific data mapping tool.
- How it will be better detailed in the next chapter, in this phase there are two activities plying a crucial role:
  - **Entity recognition & matching**: to find the real world entity within the dataset collected, and disambiguate different representations of the same real world entity.
  - **Entity mapping**: to map such entities with the KG's layer produced in the previous phase.

iTelos - Stratified approach (1)

- The iTelos phases are structured following the **stratified approach**, where the data heterogeneity is handled at different levels, to be turned into **Diversity**.
  - Information, Language, Schema and Data values.
- The stratification is present also at **process level**, where the different phases aims at handling a specific diversity layer.

iTelos - Stratified approach (2)

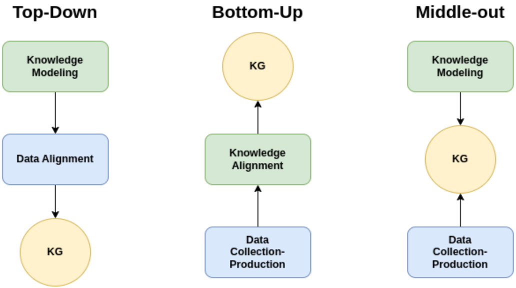
- To be more precise, the phases dedicated to the language and schema layers, take in input the result of the execution of other two layer specific application of iTelos [33]:
  - **Language (L)Telos**, producing reference standard languages, and;
  - **Knowledge (K)Telos**, producing reference standard ontologies.



[33] for this reason the top level methodology is also called Data (D)Telos

iTelos - Middle-out approach

- iTelos builds KGs adopting a **middle-out approach** between knowledge and data, so that it is
  - not too much focused on the knowledge layer (top-down approach), thus causing **hard data adaptation**;
  - neither too much focused on the data layer (bottom-up approach), thus causing **hard knowledge modeling**.



iTelos - Evaluation

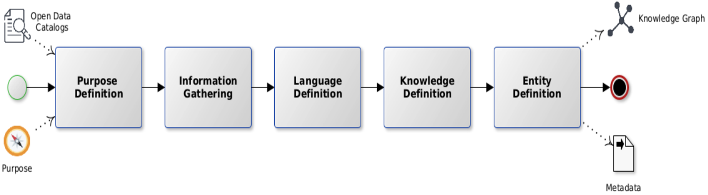
- At the end of each iTelos phase, an **evaluation activity** exists to verify that the phase output is good enough for proceeding to the next phase.[34]

- if that is not the case the process **goes backward to the evaluation activity of the previous phase**.
    - In this way the process can potentially goes backward **until the beginning**, thus allowing the user to review all the output of all the phases, and eventually do that again.

[34] See iTelos full structure

iTelos - Metadata

- For each iTelos phase, it exists an activity of **metadata definition**.
  - such activities define a parallel process which aims at producing metadata for the different resources composing the KG(s). [35]
    - In this way iTelos enables the **distribution of high quality data**.



[35] See iTelos full structure

iTelos - Data distribution

- To enable the circular data economy, iTelos, as part of the methodology, provides the possibility to publish (with different security and protection levels) the high-quality data produced.
- The iTelos output(s) are distributed by a dedicated data distribution environment called **LiveData**.
- LiveData is a **world-wide data distribution network** composed by different nodes, where each node includes a **data catalog** dedicated to the distribution of the data produced by iTelos.
- A node is created, and connected to the network, when a new **organization, or even single users**, start to adopt iTelos to produce context specific high-quality KGs.
    - Example of LiveData nodes:
        - **LiveData UniTN**: the node for the data about the University of Trento.
        - **LiveData NUM**: the node for the data about the National University of Mongolia.

iTelos - LiveData

- Moreover, LiveData includes two centralized data catalogs for the collection of standard reference languages and ontologies. Such a catalogs are called, **LiveLanguage** and **LiveKnowledge**, respectively.
- The LiveData network is accessible by the main LiveData portal: LiveData