Part 4
The iTelos Methodology

Part 4.1
A Methodology for Data Reuse

A Methodology for Data Reuse

- **iTelos** is a methodology for data reuse, which **defines the structure of the data reuse processes**.
  - Let's recap and highlight some important key points!
- **iTelos** is the methodology adopted by the **Data Intermediary**.
- **iTelos** supports **EML**.
- **iTelos** processes are based on the **DSDM architecture**.

A Methodology for Data Reuse

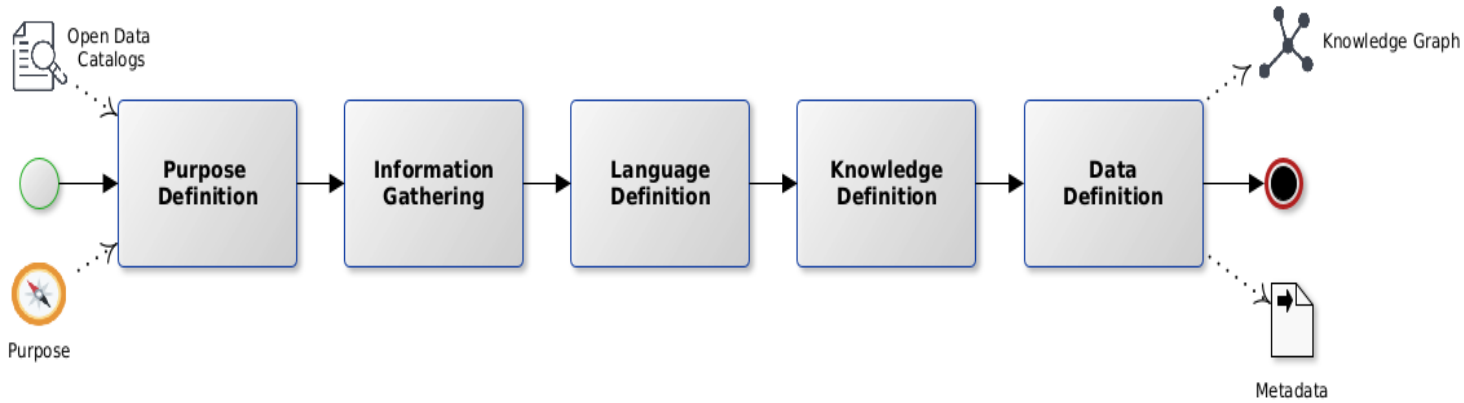
- **1 structure**: phase-based structured methodology.
  - **3 processes**: LTelos, KTelos, DTelos [30]
    - **2 roles**: data reuse for producer, and for consumer

[30] KGE is mainly focused on DTelos!

iTelos Methodology - The Structure

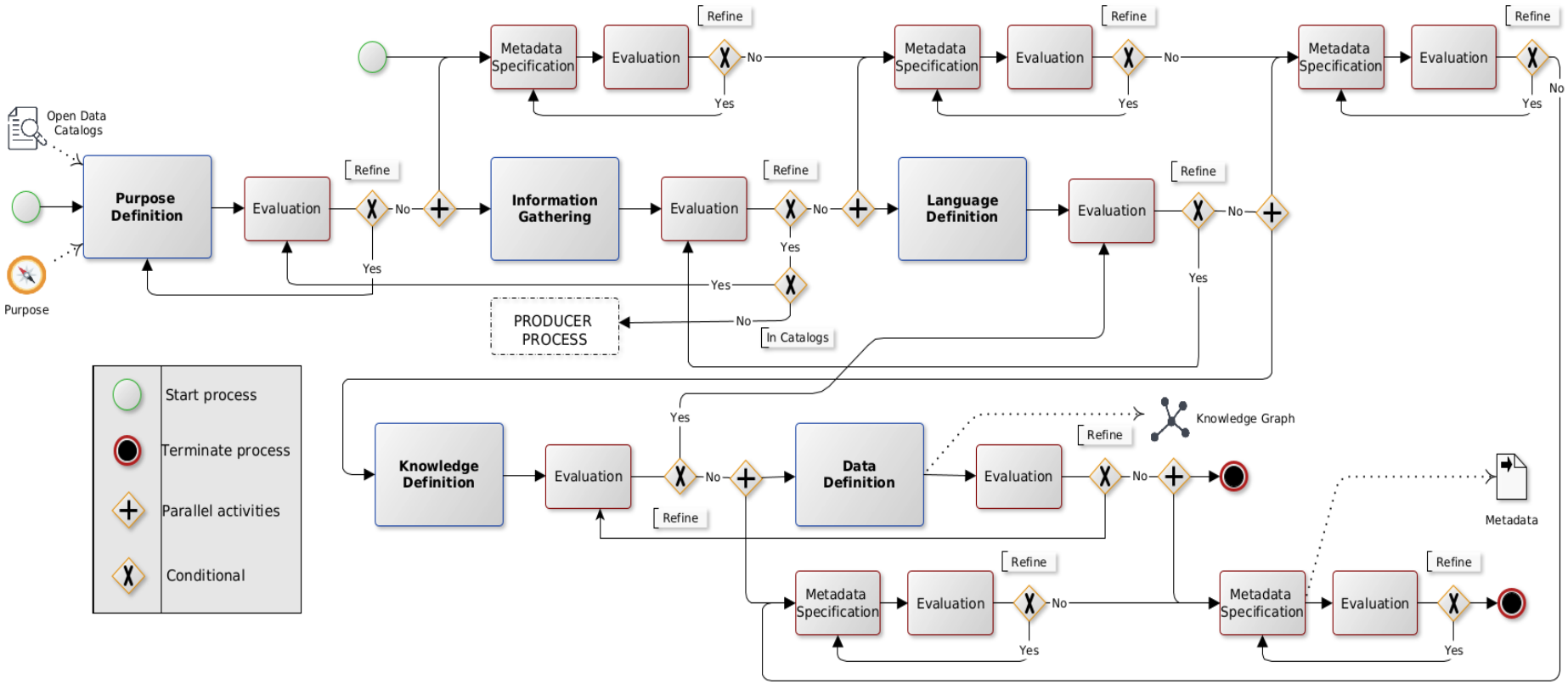
iTleos is structured in 5 well-defined phases, summarized here below [31]

- Purpose Definition (PD)
- Information Gathering (IG)
- Language Definition (LD)
- Knowledge Definition (KD)
- Data Definition (DD)



[31] the phases will be detailed in Part 4 - iTelos methodology

iTelos Methodology - The Full Structure



Purpose Definition

- Both data producer and consumer consider their own objective when building KGs.
- Such an objective implicitly includes the user **"point of view", the interpretation that the user has, and uses, to model (a portion of) the world**, where the desired information lives.
- We define the user objective as **"The Purpose"** which will lead the entire KGE process.

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 stratification layer**: data, knowledge and language.
- Notice how, depending by the agent that executes the process, the resources collected have different levels of quality:
  - intermediary producer: the resources are collected from the "disordered world", thus the quality level is, in average, lower.
  - intermediary consumer: the resources are collected from the "ordered world", 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 ontology 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.

Data 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.

Data Definition

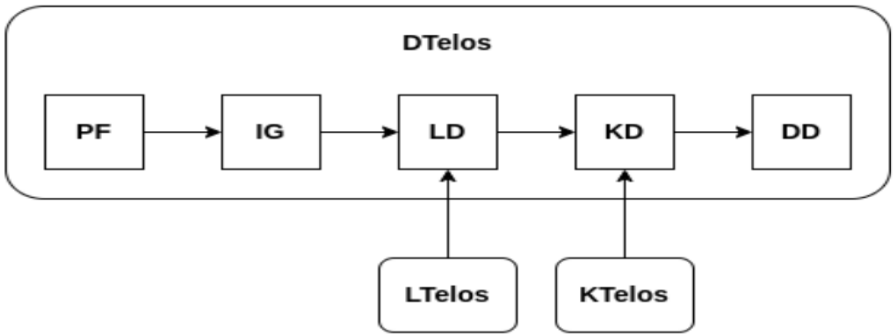
- The last phase of the methodology is dedicated to the data layer of the final KG(s).
- It is supported by a specific data mapping tool.
- How it will be better detailed in the next chapter, in this phase there two activities plying a crucial role:
  - Entity recognition: to find the real world entity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

General methodology characteristics - Producer & Consumer

Lecture (Week)	Producer	Process	Consumer	Notes
1	**Requirement Def:** Type of data sources considered and type of output to be produced ER model for each resource	**Purpose definition**	**Requirement Def:** Purpose Definition Definition of Functional and Non-Functional Requirements Purpose ER model	
2	Informal **data** and **knowledge** collection-extraction-creation	**Information Gathering**	Collection of **knowledge** and **data** form the Intermediary data mesh	Notice the difference in terms of effort, between extract non quality data and quality data from the mesh
3	**Concept** Identification for each single dataset to be produced. Single datasets UKC alignment **Dataset** filtering	**Language Definition**	Purpose specific **concept** alignment with the UKC. **Dataset** filtering	
4	Selection and alignment of **Teleology** and **Teleontology** for each single dataset **Dataset** Cleaning and Formatting (Standard, format, vocabularies and data types)	**Knowledge Definition**	Purpose specific **Teleology** and **Teleontology** selection and alignment + **Dataset** Alignment (With the teleontology)	Notice how, in the producer, the data layer requires a strong effort while the effort for the knowledge layer is reduced. And the opposite situation happens in the consumer.
5	Construction of **KGs** for each dataset	**Data Definition**	Construction of a single **purpose-specific KG**	

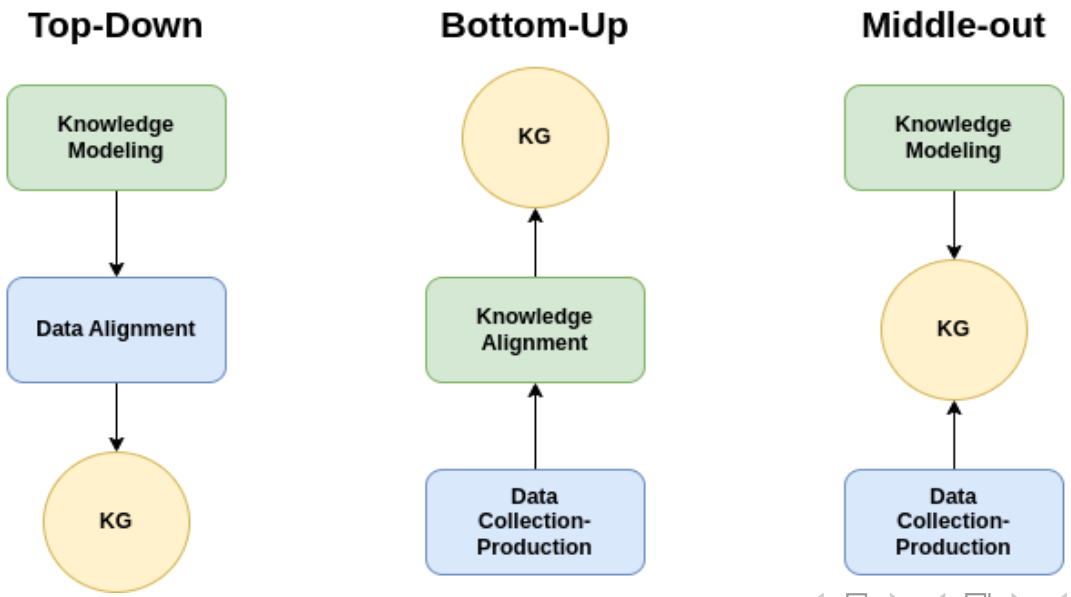
General methodology characteristics - Stratification

- The iTelos phases are structured following the **stratified approach**.
- Along the process the heterogeneity is handled at different level:
  - Source, Format, Structure, Meaning.
  - Language, Knowledge and Data.
- The stratification is present also at **process level**, where a top-level process **DTelos** aims at producing KGs by taking in input Language and Knowledge resources produced by other two processes, **LTelos** and **KTelos** respectively.



General methodology characteristics - 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**.



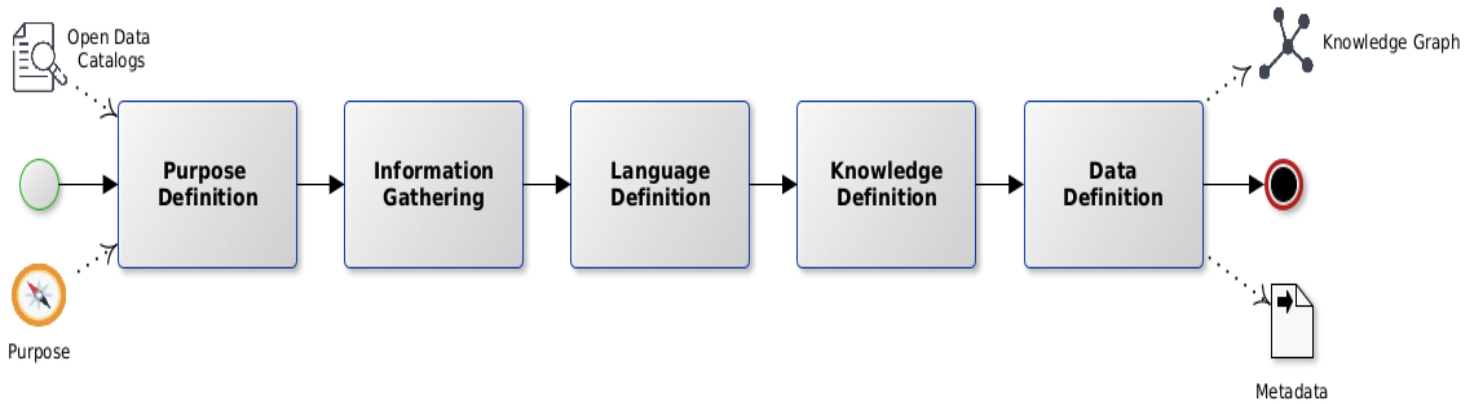
General methodology characteristics - Evaluation

- 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.[33]
- 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.

[33] See iTelos full structure

General methodology characteristics - 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). [34]
    - In this way iTelos enables the **distribution of high quality data**.



[34] See iTelos full structure

iTelos Project set-up

- Each iTelos project needs a specific **repository**, where the resources (Language, Knowledge, Data and Metadata) are maintained during the process execution.
  - such a repository can be cloned by a github template repository.
- **Documentation** is a crucial part during the execution of the iTelos process.
- A **project report** has to be completed at the end of the process execution.
  - At the end of each phase, a report template document has to be filled, by reporting the execution of the current phase activities.
- At the end of the project, a **set of slides summarizing the work done** needs to be produced and stored in the repository together with the project report, into the dedicated "Documentation" directory.

iTelos Project publication

- In order to be properly published in the KGE project catalog, a static web page of the whole project has to be produced.
  - It can be done quickly by creating a github page directly from the project repository.
- Such a web page, will be directly linked in the catalog, thus allowing the users to look for the quality resources produced by the project.