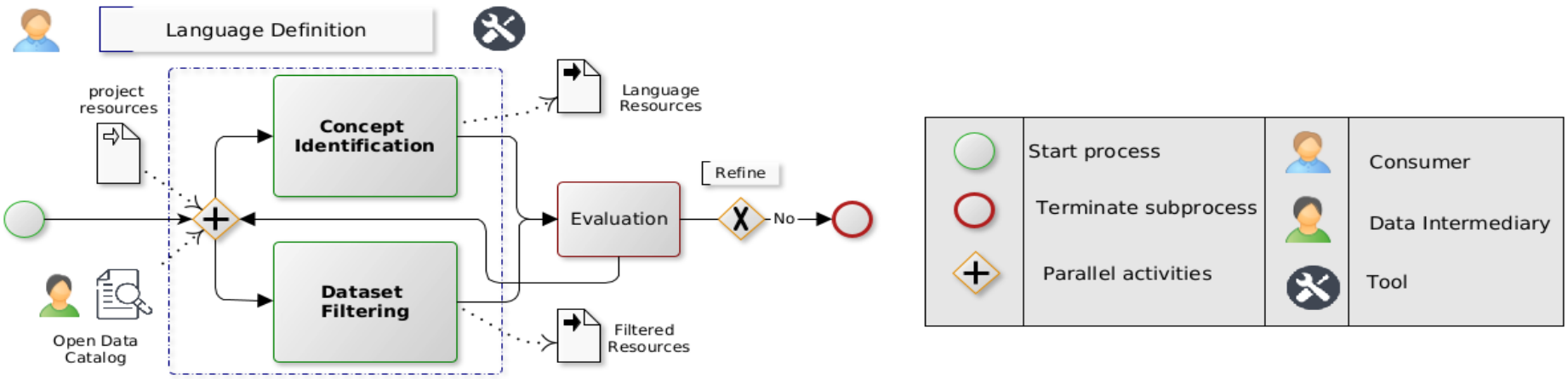
Part 5.5
Phase 3 - Language Definition

iTelos Language Definition Phase

Phase 3 - Language Definition - Structure



- **Input**: Purpose Formalization sheet, ER model, Formalized resource set.
- **Objective**: Formally define the concepts used to represent the information included in the final KG.
- **Output**: Language resources (Formal concept definition), Filtered resource set.

Phase 3 - Language Definition

Activity 1 - Concept Identification

- From the above, we have the notions of:
  - which are the elements (concept structure) used to define a language;
  - which are the knowledge bases available to maintain such elements;
  - how to identify uniquely a concept;
  - what is a natural language and a domain language.
- How this notions are considered by the iTelos methodology ?
  - By exploiting a specific **process of concept identification** [40].

[40] **Note**: such a process is part of the LTelos process producing language resources.

Activity 1 - Concept Identification

- The activity process aims at **defining the purpose-specific language resources** for the current iTelos execution.
- The process is composed by the following steps:
  1. Select the purpose-specific concepts to be formalized.
  2. Check if the concepts have been already defined in the UKC.
     1. If yes, collect the formal concepts definitions.
     2. If no, define the new concepts formally.
  3. Build the purpose-specific language file including the above formal concepts definitions.

Step 1 - Concepts Selection

- The objective of the first step is to **select all the concepts** to be used to represent the information in the final KG.
- Such concepts are those representing:
  - **ETypes**
  - **Data** and **object properties**
  - Eventually, specific concepts used as data properties values.
- Due to that, the concepts can be selected from the resources produced in the previous iTelos phases.
  - From the purpose **ER model and PFSheet**.
  - From the Language, Data and Knowledge **resources collected**.

Step 2 - UKC alignment [41]

- The objective of the second step is to **find**, or **define**, the **formal definition for each concept** selected before.
- To this end, the UKC is exploited, where several concepts are already defined.
- The key idea is that,
  - if a concepts to be formalized is already present in the UKC, we will **get the formal definition** from the UKC itself;
  - if, instead, such a concept is not present in the UKC, **it will be defined formally, and later eventually uploaded in the UKC** (quality check is required), for further reused.

[41] A practical lecture with a dedicated tool will show how to concretely execute this step.

Step 2 - UKC alignment - Identification

- The formal definition for a concept is composed as follows:
  - ConcpetLabel _UKCIndetifier
  - Example: Hospital_GID-10045
- The **UKCIdentifier** is a numeric value within a range. Such a **range defines the UKC ID's space for all the concepts of a specific purpose**. Each range is associated to a purpose-specific **XML namespace**.

Step 2 - UKC alignment [Notes]

- **Note 1**: The number of purpose-specific concepts to be formally defined in this steps, depends on how many concepts for the purpose's domain, have been uploaded in the UKC (reference domain standard vocabularies).
- **Note 2**: The concepts categorized as Common have more probability to be found in the UKC, while for Core and Contextual concepts the probability decrease, thus requiring more effort in concept formalization.
- **Note 3**: An increasing adoption of the iTelos methodology implies an increasing number of concepts added in the UKC, for different domains, thus actually reducing the concept formalization effort over time.

Step 3 - Language resource building

- The final step of the Concept Identification process, aims at **generating the file representing the language resources** for the purpose considered into the relative iTelos execution.

- To this end, the concepts, formally defined in the previous steps, are collected into a **spreadsheet** having at least three columns:

  - the first column list down the **concept IDs** fro the concepts defined;
  - the second column lists down all the formal **concepts labels, or words**, and;
  - the third column provides a **description** (called "gloss" in the UKC) of the meaning for each relative concept in the first column.

- Additionally, more columns can be added to specify **the words and their relative gloss in other languages**.

Language resource & Language Teleontology

- The concepts listed into the Language resource created for your projects, represent the set of terms (or words) of your purpose-specific domain language.
- This means that this concepts can be used to define a Language Teleontology (LT) for your specific purpose.
  - For lack of time, this course will not cover the generation of the LT.

Phase 3 - Language Definition

Activity 2 - Dataset Filtering

- The Data set Filtering is the activity of the current iTelos phase, focused on the final KG's **data layer**.
- This activity aims at **aligning the data layer resources**, previously collected and formalized, **with the concepts** identified and formalized in the parallel knowledge layer activity.
- Concretely, this activity **filters out**, from the current resource set, all **the elements** (entities, attributes, ETypes and properties) which are **not defined by any of the concepts formalized** in the parallel Concept Identification activity.

Phase 3 - Language Definition - Summary

- What has been done in this phase.
- The **heterogeneity at language level** has been handled.
  - By defining a **purpose-specific domain language** (thus based on a natural language),
  - composed by **concepts formally defined and uniquely identified** (associated to a purpose-specific namespace).
- The **purpose-specific language resource** for the final KG has been created.
- The **data** resources have been **filtered and aligned with the language's concepts** defined for the final KG.

Phase 3 - Language Definition

- **Language Definition - Short Demo**