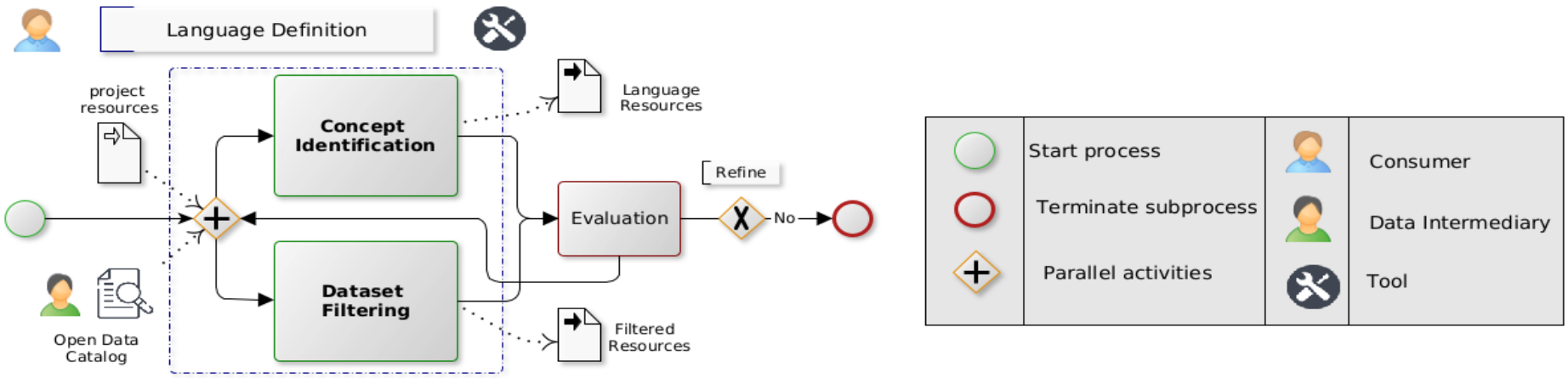
Part 4.4
Phase 3 - Language Definition

Phase 3 - Language Definition



- **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.

Language Definition - Activities

- In this phase, like in the others, the activities are divided over the **knowledge** and **data** layers.
- The objective of this phase is to:
  - (knowledge layer) **identify and formalize** the "language elements", or more formally **the language concepts**, used to represent the information to be included in the final KG;
  - (data layer) **filter the resources collected** (both knowledge and data) **based on the concept set** identified and formalized.
- In other words, in this phase iTelos aims at formally defining the **language of the KG**.
  - To do that, **iTelos reuses as much as possible the concepts from already existing natural and domain languages**.

Language Definition - Activities

- (knowledge layer) **Concept Identification**: This activity aims to formally identify and define the concepts to be used for the representation of ETypes and Properties, into the final KG.
  - To this end **the data layer** (datasets values) **has to be considered** to better understand the meaning (concept definition) of each concept to be identified.
  - This activity produces a specific output including the language resources (concepts) identified.
- (data layer) **Dataset Filtering**: This activity aims at filtering out the entities and attributes which are not described by at least one concept identified in the knowledge layer activity.

Language Definition - Producer & Consumer

- **Producer**: at producer side the objective is to identify the concepts for the ETypes and properties included in each single formal resources to be produced.
  - This means that **more language resources files are produced**, on for each KG generated by the Producer.
- **Consumer**: at consumer side the objective is to identify the concepts for the ETypes and properties included in the (single) final KG.
  - In this case a **single language resource file is produced**.

Phase 3 - Language Definition

Preliminaries definitions

- To achieve the intermediate phase's output it is important to understand how such output has to be represented.
- The concepts, as elements of the KG's language, have to be formally defined following a specific structure and formats.
- The preliminaries definitions describe how the **language resources** have to be properly defined.

Preliminaries definitions - Language structure

- **Synonymy:** A word with the same (or nearly the same) meaning (i.e., *sense*) as another word. e.g., car, auto, automobile, etc.
- **Polysemy:** The coexistence of many possible meanings for a word. e.g., 645 distinct meanings of the word *run*.
- **Synset:** A synset is a set of synonyms that are, in principle, interchangeable for a particular sense of a word. e.g., {car, auto, automobile, motorcar}.
- **Subsumption:** A classification of concepts from the general (i.e., *hypernym*) to the specific (i.e., *hyponym*) via `IS-A` relation. e.g., spoon `IS-A` cutlery.
- **Lexical Gap:** The absence of a word in a particular language where it is present in another. e.g., *Malga* in Italian absent in English.

Phase 3 - Language Definition

Wordnet and UKC

- To support the concept definition, as defined above, an environment (knowledge base) is required to support such definitions.
- The environment considered is the **Universal Knowledge Core (UKC)** project based on the **WordeNet** project.
  - (While WordNet is more oriented on Natural Languages (NL), the UKC is exploited for Domain Languages (DL) too)[39]
- Here below they are both described.

[39] See next slides for more details about natural and domain languages.

WordNet

- WordNet is a large lexical database of English. It is hosted by the Princeton University.
- WordNet **interlinks** not just word forms (e.g., nouns) — strings of letters — but specific ***senses*** of words into *synsets* each expressing a distinct concept.
- A synset in WordNet contains a brief definition (***gloss***) and, in most cases, one or more short sentences illustrating the use of the synset members. Word forms with several distinct meanings are represented in as many distinct synsets.
- The most frequently encoded relation among synsets is the super-subordinate relation (also called hyperonymy, hyponymy or **IS-A relation**).
- The resulting network of WordNet is a **network of meaningfully related words and concepts**.

WordNet: Illustration

WordNet Search - 3.1

- WordNet home page - Glossary - Help

Word to search for: car [Search WordNet]

Display Options: (Select option to change) [Change]

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

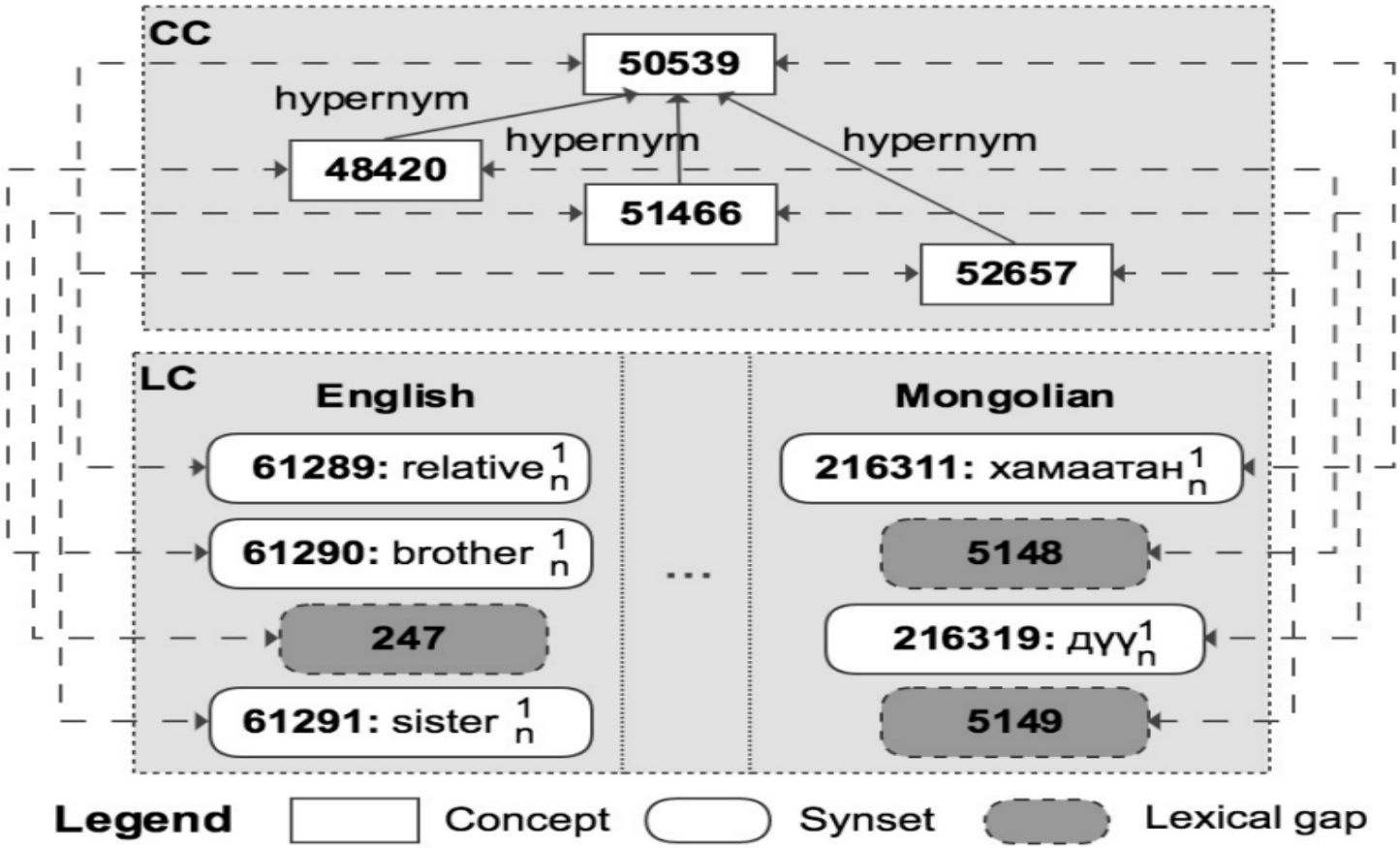
- S: (n) **car**, auto, automobile, machine, motorcar (a motor vehicle with four wheels; usually propelled by an internal combustion engine) *"he needs a car to get to work"*
- S: (n) **car**, railcar, railway car, railroad car (a wheeled vehicle adapted to the rails of railroad) *"three cars had jumped the rails"*
- S: (n) **car**, gondola (the compartment that is suspended from an airship and that carries personnel and the cargo and the power plant)
- S: (n) **car**, elevator car (where passengers ride up and down) *"the car was on the top floor"*
- S: (n) cable car, **car** (a conveyance for passengers or freight on a cable railway) *"they took a cable car to the top of the mountain"*

UKC - The Reference Ontology

- Universal Knowledge Core (UKC) (see: CICLing paper) is a reference lexical-semantic ontology (developed by the Knowdive[40] research group) to which all DLs can be hierarchically aligned.
- The UKC is formally structures a knowledge base composed by two main components:
  1. **Language Core (LC)**: focused on modelling WordNet-like NL and DL hierarchies
  2. **Concept Core (CC)**: focused on integrating LC hierarchies in a language-independent semantic hierarchy.

[40] KnowDive, University of Trento, DISI department

UKC Illustration



UKC [Contd.]

- **LC** is comprised of NL and DL hierarchies (e.g., English and Mongolian in the example). These hierarchies are directed acyclic graphs (see: DAGs) of language-specific *synsets* (sets of synonyms) (e.g., *relative*, *brother*, *sister*) structured via subsumption (`IS-A`) relationship in an overall hierarchy.
- **CC** is a **language-independent** semantic hierarchy whose nodes are language-independent abstractions of semantically synonymous language-specific synsets in LC. CC nodes have a **unique identifier** (named GID) (e.g., *50539* for the synset *relative*).

Application example of UKC

Phase 3 - Language Definition

XML Namespaces

- In this specific phase of the methodology we need to identify uniquely the concepts.
- Once identified they can be represented and maintained in the UKC as described above.
- The set of concepts identified, and formally represented, is a **Language resources** (that has been produced by a specific project/purpose).
  - Language resources are one of the main types of resources produced (and distributed) by iTelos.
- **How to identify uniquely a concept ?**
  - It can be done by using the **XML Namespaces**.

What is XML?

- XML stands for eXtensible Markup Language
- XML is a markup language much like HTML
- XML was designed to store and transport data
- XML is a W3C Recommendation
- All major browsers have a built-in XML parser to access and manipulate XML.

XML vs. HTML

The Adventures of Tom Sawyer

Front piece of *The Adventures of Tom Sawyer*

Author	Mark Twain
Cover artist	created by Mark Twain
Country	United States
Language	English, Limited Edition(Spanish)
Genre	Bildungsroman, picaresque, satire, folk, children's novel
Publisher	American Publishing Company
Publication date	1876[1]
OCLC	47052486
Dewey Decimal	Fic. 22
LC Class	PZ7.T88 Ad 2001
Followed by	*Adventures of Huckleberry Finn*

HTML: focus on presentation

```
<h2>The adventures of Tom Sawyer</h2>
...
<b>Author: </b> Mark Twain <br>
<b>Cover artist: </b> created by <a href="http://...">Mark Twain </a>
...
```

XML: focus on metadata

```
<book>
        <title> The adventures of Tom Sawyer </title>
        <author> Mark Twain </author>
        <genre> Bildungsroman </genre>
        <genre> picaresque </genre>
        ...
        <publisher> American Publishing Company </publisher>
        <year>1876</year>
</book>
```

XML Namespaces

- XML *namespaces* provide a method for qualifying concept names used in XML documents by associating them with namespaces identified by URI references.
- The Prefix provides the *namespace prefix* part of the qualified name, and MUST be associated with a namespace URI reference in a namespace declaration.
- e.g., *foaf* as the namespace prefix for *http://xmlns.com/foaf/0.1/*
- However, XML documents, often, exhibit **semantic heterogeneity and ambiguity** when described even with lexically similar names from different XML markup vocabularies codified as namespaces.

A Fragment of FOAF Namespace

[back to top]

Class: foaf:Document

Document - A document.

Status: testing

in-range-of: foaf:accountServiceHomepage foaf:publications foaf:workplaceHomepage foaf:page foaf:workInfoHomepage foaf:homepage foaf:schoolHomepage foaf:weblog foaf:tipjar foaf:interest

in-domain-of: foaf:sha1 foaf:topic foaf:primaryTopic

The foaf:Document class represents those things which are, broadly conceived, 'documents'.

The foaf:Image class is a sub-class of foaf:Document, since all images are documents.

We do not (currently) distinguish precisely between physical and electronic documents, or between copies of a work and the abstraction those copies embody. The relationship between documents and their byte-stream representation needs clarification (see foaf:sha1 for related issues).

[back to top]

Class: foaf:Group

Group - A class of Agents.

Status: unstable

in-domain-of: foaf:membershipClass foaf:member

The foaf:Group class represents a collection of individual agents (and may itself play the role of a foaf:Agent, ie. something that can perform actions).

This concept is intentionally quite broad, covering informal and ad-hoc groups, long-lived communities, organizational groups within a workplace, etc. Some such groups may have associated characteristics which could be captured in RDF (perhaps a foaf:homepage, foaf:name, mailing list etc.).

While a foaf:Group has the characteristics of a foaf:Agent, it is also associated with a number of other foaf:Agents (typically people) who constitute the foaf:Group. FOAF provides a mechanism, the foaf:membershipClass property, which relates a foaf:Group to a sub-class of the class foaf:Agent who are members of the group. This is a little complicated, but allows us to make group membership rules explicit.

Phase 3 - Language Definition

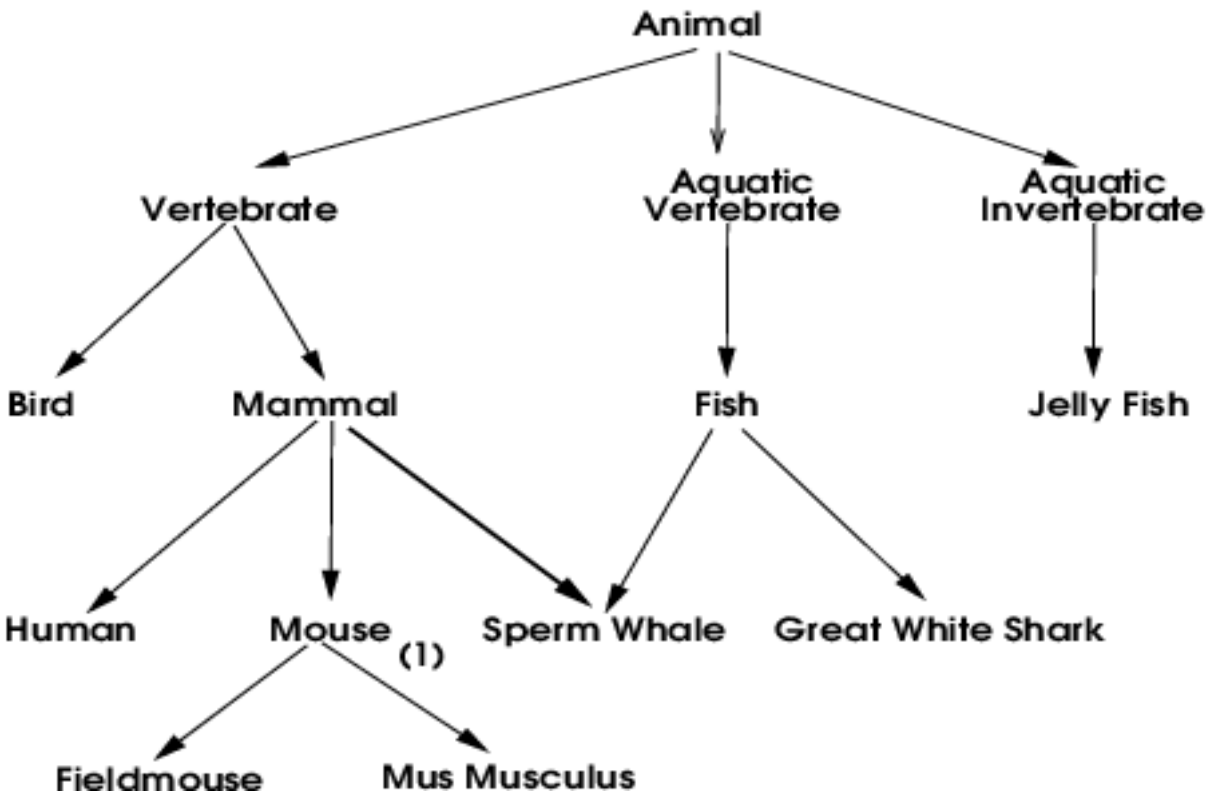
Natural and Domain languages

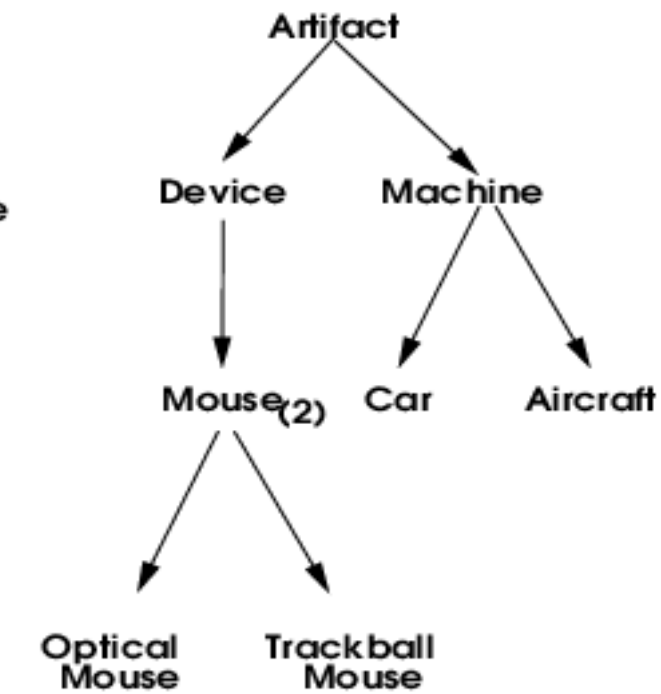
- As already anticipated, concepts can be defined for **Natural Languages (NLs)** and **Domain Languages (DLs)**.
- How can we distinguish between them ?

Natural Language

- A **natural language** (NL) is any *"language that **occurs naturally** in a human community by a process of use, repetition, and change **without** conscious planning or premeditation."* - Wikipedia
- Examples: ***Italian***, English, etc.
- NLs reflect ***Language Diversity*** rooted in genetic ancestry, geography, culutre within and across NLs (see: IJCAI paper).
- NLs also encode **semantic ambiguity** in terms of different linguistic phenomenon, e.g., (see: CICLing paper):
  - *synonyms*, e.g., car, auto, automobile, motorcar, etc.
  - *polysemy*, e.g., 645 distinct meanings of the word *"run"* (see: NPR).
- NLs are computationally formalized as Word-Net like **lexical-semantic hierarchies** (see: WordNet).

WordNet English: Example





Domain Languages

- A **domain language** (DL) is a language that is **artificially created and curated** by a human community as ***a controlled vocabulary*** for use and repetition **with** conscious planning and premeditation.
- A DL is defined by taking a base NL (e.g., English) and enriching it with **uniquely-identified domain-specific words** (e.g., English enriched with Healthcare facilities terminology).
- **Communities** are crucial to DLs, e.g., a healthcare community curated-DL would encode a wider and fine-grained coverage of healthcare facilities than a general facilities-DL.
- DL terms have **precise and unambiguous semantics**, thereby, solving the ambiguity of both NLs and XML namespaces.
- DLs, like NLs, are computationally formalized as **lexical-semantic hierarchies**.

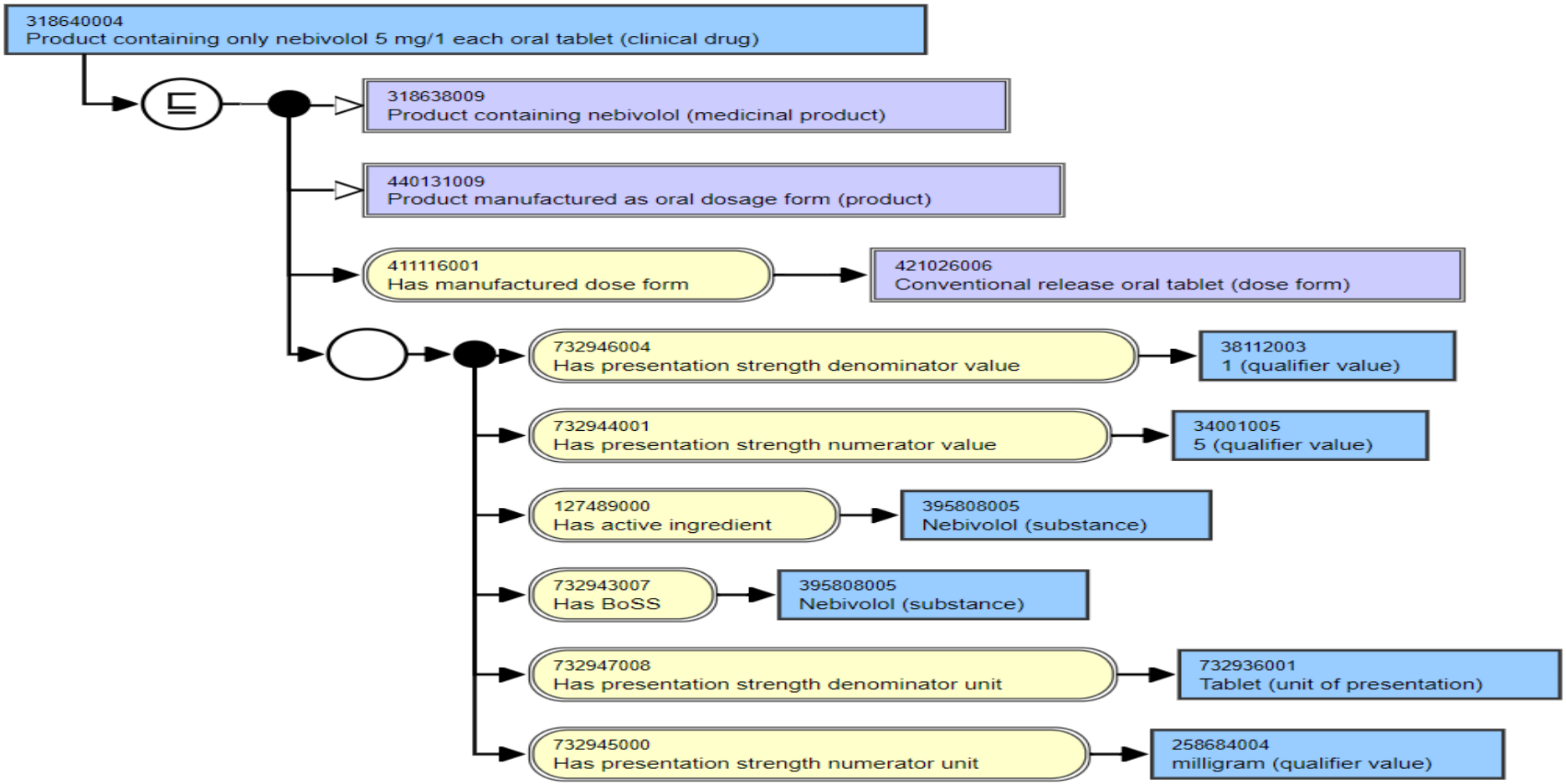
DL in English: Example

concept labels	description
point_bar_GID-10000	a bar as a point place that offer drinks.
point_bus_station_GID-10001	a large bus station as a point place with multiple platforms.
point_bus_stop_GID-10002	a bus stop as a point place.
point_cafe_GID-10003	a cafe as a point place.
point_library_GID-10004	a library as a point place.
point_peak_GID-10005	a mountain peak as a point place.
point_railway_halt_GID-10006	a smaller, local railway station, or subway station as a point place.
point_railway_station_GID-10007	a larger railway station as a point place of mainline rail services.
point_restaurant_GID-10008	a normal restaurant as a point place.
point_spring_GID-10009	a spring as a point place, possibly source of a stream.
point_supermarket_GID-10010	a supermarket as a point place.
point_tree_GID-10011	a tree as a point place.
primary_GID-10012	a primary road, typically national.
secondary_GID-10013	a secondary road, typically regional.
step_GID-23917	support consisting of a place to rest the foot while ascending or descending a stairway
tertiary_GID-10014	a tertiary road, typically local.
very_small_road_service_GID-10015	a service road for access to buildings, parking lots, etc.
pedestrian_GID-10016	a pedestrian only street.
bridleway_GID-10017	a path for horse riding.
cycleway_GID-10018	a path for cycling.

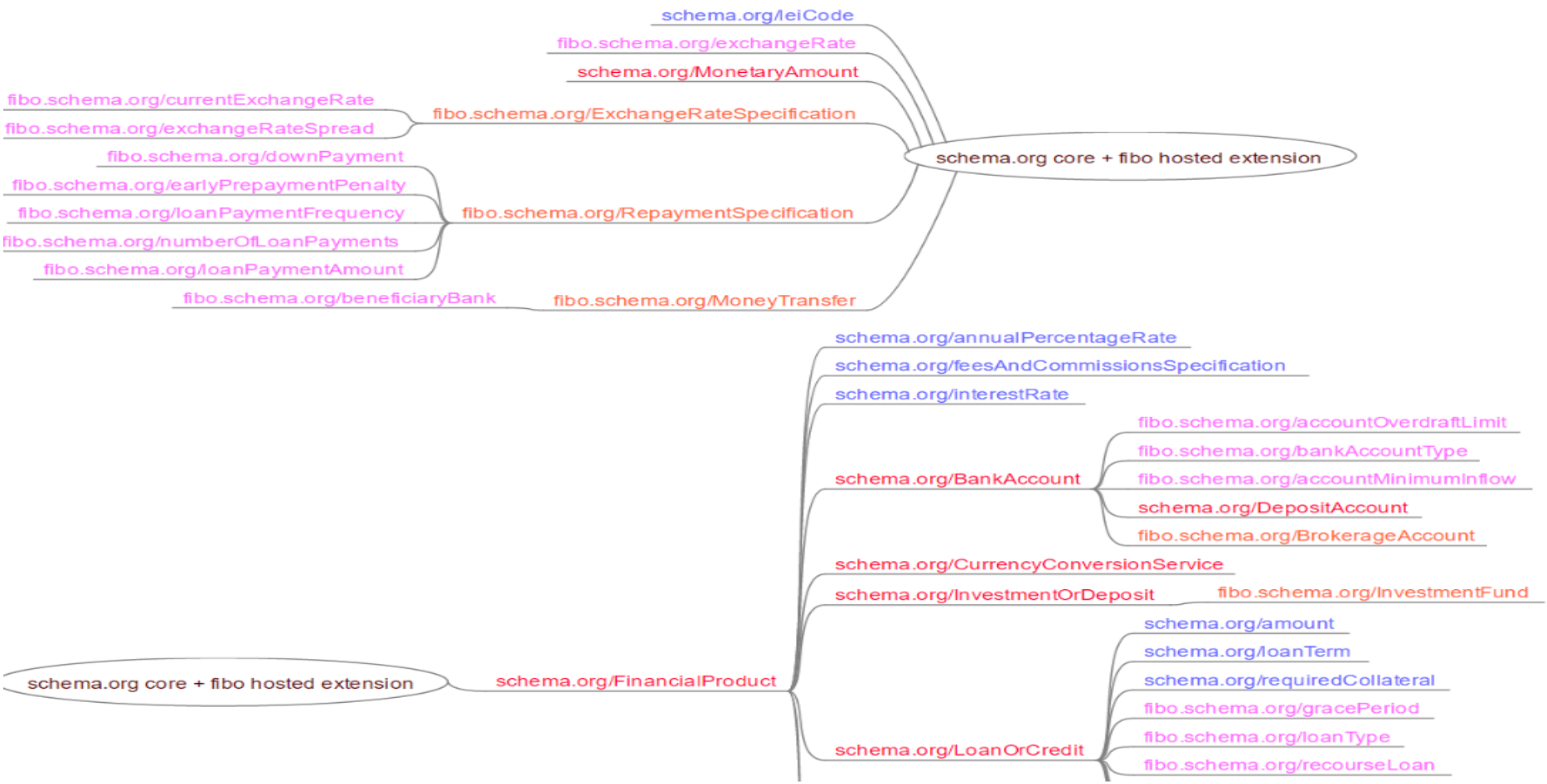
Why do we need Domain Languages?

- Domain Languages are crucial to Knowledge Graph Engineering in two aspects:
  - as ***Language of Data***;
  - to ensure ***Linguistic Interoperability***.
- DLs are the only possible means to annotate and describe datasets, i.e., what we define as the ***language of data*** (see: COLING paper).
- When two datasets are annotated and described using the same DL, they become mutually ***interoperable*** in terms of the language (both NL and DL) in which they are defined.
- This **absorbs syntactic heterogeneity** which is a major impediment to integrate data and knowledge at large scale.
- DLs are easily exploitable by domain-driven NLP techniques and applications (ECAI Workshop Paper - pag 49)

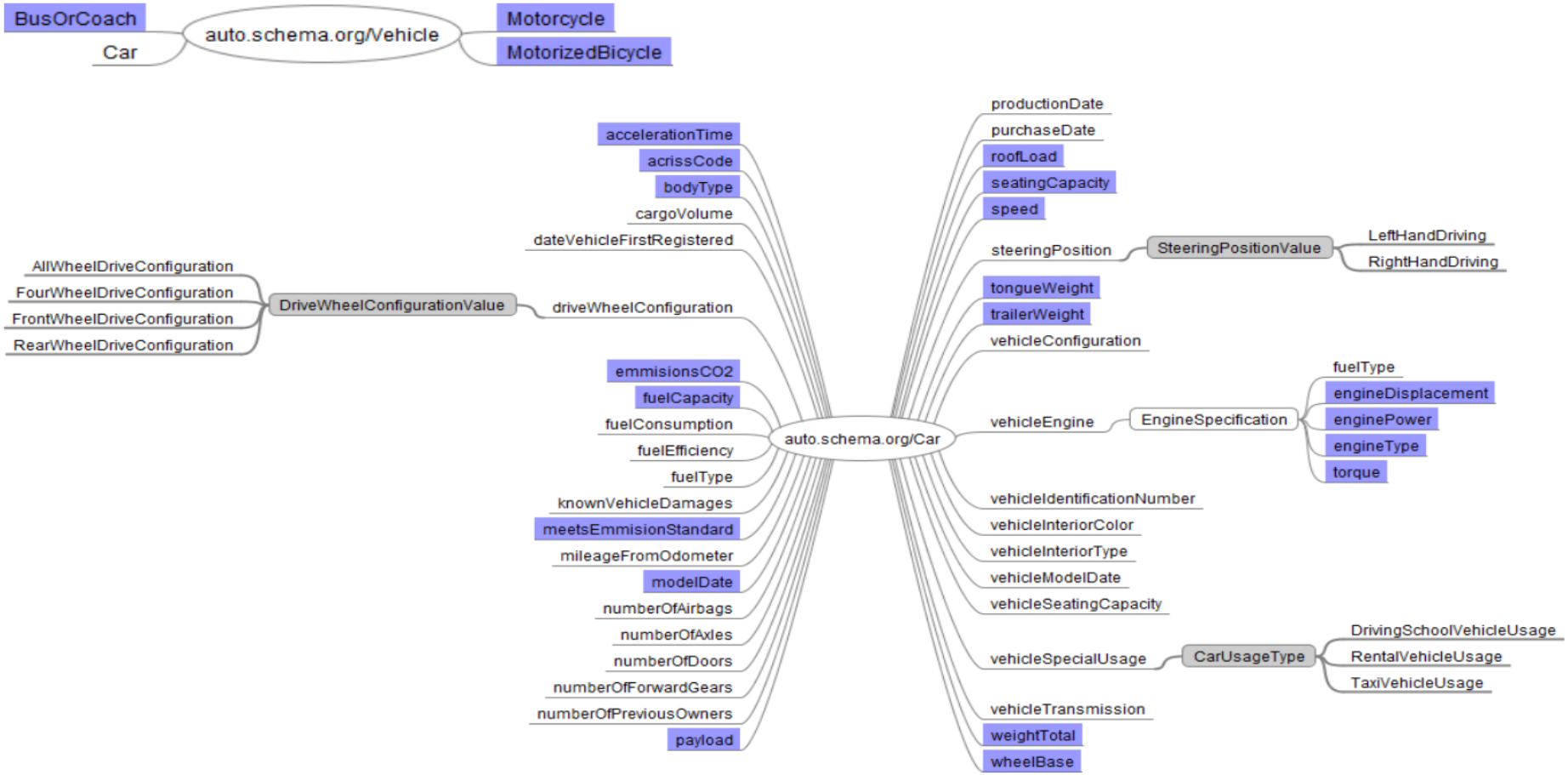
Example Healthcare DL



Example Banking and Finance DL



Example Automotive DL



DLs generation

- There are several dimensions to be factored in for creating a domain language. They are as follows:

  1. Fixing the base Natural Language
  2. Fixing the domain which the DL should describe
  3. Fixing the domain-specific terminology within a domain which the DL should describe (e.g., in a structured tabular format)
  4. Fixing the reference lexical-semantic ontology (e.g., WordNet) which provides hierarchical structure to the DL
  5. Aligning and integrating the DL to the reference lexical-semantic ontology.

- **Note:** Notice that, preferrably, in addition to the above dimensions, a publicly accessible data catalog should be instantiated for browsing and downloading DLs for reuse.

- **Note 2:** The DataScientia catalogs offer a set of language resources already developed for several different NLs.
  - The concepts within such language resources **can be reused to compose the DL for your specific Purpose**.

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** [41].

[41] **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**
- 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 Data and Knowledge **resources collected**.

Step 2 - UKC alignment [42]

- 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.

[42] 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.

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 concept formally defined in the previous steps, are collected into a **spreadsheet** having two columns:
  - the first column lists all the formal **concepts labels**, and;
  - the second column provides a **description** (called "gloss" in the UKC) of the meaning for each relative concept in the first column.

concept labels	description
bar_GID-14950	a room or establishment where alcoholic drinks are served over a counter
bus station_GID-15745	a terminal that serves bus passengers
cafe_GID-15804	a small restaurant where drinks and snacks are sold
library_GID-20054	a building that houses a collection of books and other materials
restaurant_GID-22500	a building where people go to eat
supermarket_GID-24168	a large self-service grocery store selling groceries and dairy products and household goods
train station_GID-22321	terminal where trains load or unload passengers or goods
id_GID-10032	short for identifier
latitude_GID-46263	the angular distance between an imaginary line around a heavenly body parallel to its equator and the equator itself
longitude_GID-46270	the angular distance between a point on any meridian and the prime meridian at Greenwich

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 **filter out**, from the current resource set, all **the elements** (entities, attribute, 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.