

Part 5.5

Phase 3 - Language Definition

- 1 KG Construction
- 2 iTelos
- 3 Phase 1 - Purpose Definition
- 4 Phase 2 - Information Gathering
- 5 Phase 3 - Language Definition**
- 6 Phase 4 - Knowledge Definition
- 7 Phase 5 - Entity Definition

Language Diversity

- 1 Language Diversity
- 2 Lexical Resources
- 3 The Universal Knowledge Core (UKC)
- 4 Language Teleontology
- 5 iTelos Language Definition Phase

Language

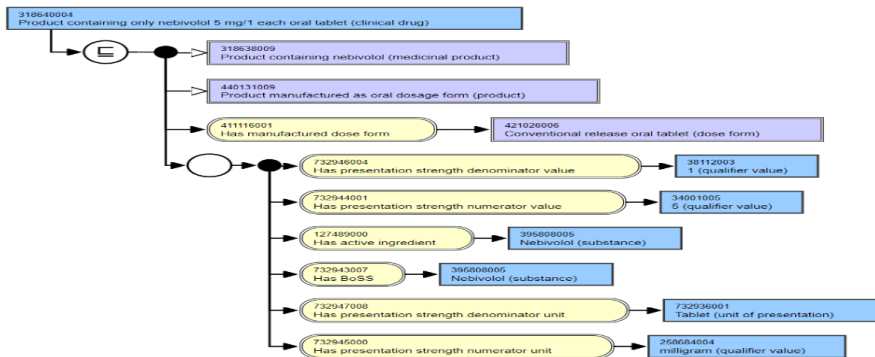
- In the context of knowledge representation via KGs, languages can be primarily of two different types:
 - Natural languages (NLs), organized via ambiguous words, but very efficient in conveying common-sense concepts, e.g., Italian;
 - Domain languages, organized via words with unique meanings, very efficient in conveying domain/application specific concepts, e.g., Italian ICD.
- We have the following levels of language diversity:
 - diversity within a natural language;
 - diversity across multiple natural languages;
 - diversity within domain languages;
 - diversity across natural and domain languages.

Domain Language

- Domain Languages (DLs) 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
- 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.

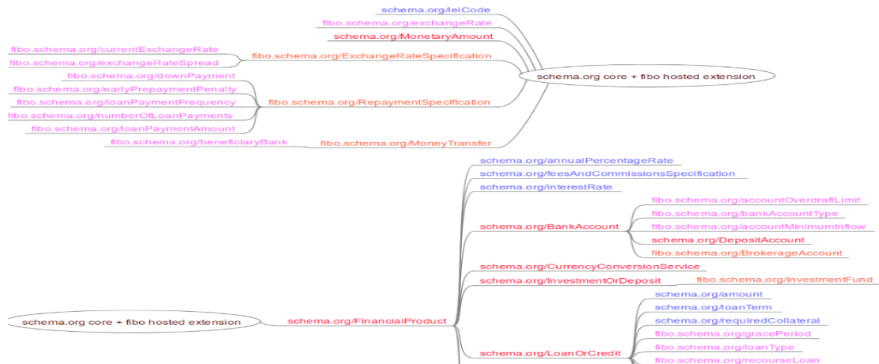
Healthcare DL - SNOMED-CT

SNOMED-CT is a healthcare DL providing codes, terms, synonyms and definitions used in clinical documentation and reporting. A fragment of SNOMED-CT focused on clinical drug is illustrated below:



Finance DL - schema.org

The financial DL of schema.org refers to the most important real world objects related to banks and financial institutions. A fragment of the finance DL is illustrated below:



Automotive DL - schema.org

The automotive DL of schema.org refers to the most important real world objects related to popular vehicles. A fragment of the automotive DL is illustrated below:



Language (Contd.)

- Linguistic phenomenon like the following allows for diverse language-level representation of entities:
 - Polysemy, the coexistence of many possible meanings (also known as *senses* for a word), e.g., bank;
 - Synonymy, words with the same or synonymous sense as another word, e.g., Big (large, huge, giant).
- Many-to-many mappings between words (which are language-dependent) and concepts (which are abstract, language-independent), both within the same language and across languages, contribute to language diversity.
- We now describe and illustrate the notions of Language Unity and Language Diversity which allow us to encode language diversity in KGs.

Language *Unity*

- 1 Language *Unity* (CU) refers to the fact that two representations are encoded using the same concept (expressed via a unique language-independent identifier), e.g., two representations of the same entity 'FP372MK' as *Car* having ID: 15944.
- 2 Our language-independent *Word Sense Disambiguation (WSD)* method [Bella, Zamboni, Giunchiglia, et al. 2016] helps in ascertaining Language *Unity*, i.e., whether two concepts are of the same sense, exploiting the language-independent concept IDs.

Language Unity: Illustration

Car LU (ID:15944)				
Nameplate	schema: speed	schema: fuelCapacity	schema: fuelType	schema: modelDate
FP372MK	150	62	Petrol	2020-11-25

Vettura LU (ID:15944)		
Targa	Velocità	Tipo di corpo
FP372MK	158	Coupé

Vehicle			
vso:VIN	vso:feature	vso:modelDate	vso:speed
FP372MK	Armrest	2020-11-25	155.0

Language *Diversity*

- 1 Language *Diversity* (CD) refers to the fact that two representations are encoded using different concepts (expressed via different language-independent senses), e.g., two representations the same vehicle 'FP372MK' as two different concepts: *Vehicle* with ID: 25142 and *Car* with ID: 15944
- 2 Our language-independent *Word Sense Disambiguation (WSD)* method [Bella, Zamboni, Giunchiglia, et al. 2016] helps in ascertaining Language *Diversity*, i.e., whether two concepts are of two different senses, exploiting language-independent concept IDs.

Language Diversity: Illustration

Car LD (ID:15944)				
Nameplate	schema: speed	schema: fuelCapacity	schema: fuelType	schema: modelDate
FP372MK	150	62	Petrol	2020-11-25

Vettura		
Targa	Velocità	Tipo di corpo
FP372MK	158	Coupé

Vehicle LD (ID:25142)			
vso:VIN	vso:feature	vso:modelDate	vso:speed
FP372MK	Armrest	2020-11-25	155.0

Lexical Resources

- 1 Language Diversity
- 2 Lexical Resources
- 3 The Universal Knowledge Core (UKC)
- 4 Language Teleontology
- 5 iTelos Language Definition Phase

Lexical Resources

- A lexical resource is defined as a language resource consisting of data regarding the lexicon of one or more natural and/or domain languages e.g., in the form of a database.
- Lexical data includes:
 - senses, as explained earlier
 - words
 - synset (set of synonymous words)
 - gloss, a natural language textual description of the sense of a word
 - examples, exemplifying the usage of a word.
- In the following slides, we describe and illustrate some important state-of-the-art lexical resources which computationally encode language unity and diversity:
 - Princeton WordNet
 - EuroWordNet
- Finally, we describe the knowdive catalogue LiveLanguage.

Princeton WordNet (PWN)

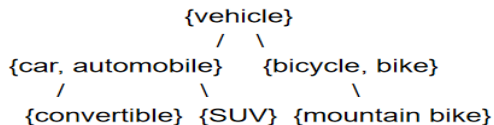
- A large lexical resource, or “electronic dictionary,” developed and maintained at Princeton University.
- Digitally encodes language diversity in terms of most English nouns, verbs, adjectives, adverbs organized as a hierarchy.
- Can be used by humans and machines
- Princeton WordNet is for English only, but it is linked to other wordnets via many other languages.
- It can be searched online here

Hypo/Hypernymy

Hypo-/hypernymy relates noun synsets

Relates more/less general concepts

Creates hierarchies, or “trees”



“A car is is a kind of vehicle” \Leftrightarrow “The class of vehicles includes cars, bikes”

Hierarchies can have up to 16 levels

Hyponymy (Example)

Hyponymy

Transitivity:

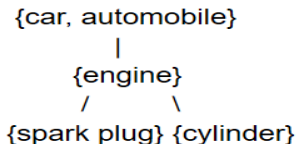
A car is a kind of vehicle

An SUV is a kind of car

=> An SUV is a kind of vehicle

Meronymy/Holonymy

Meronymy/holonymy (part-whole relation)



“An engine has spark plugs”

“Spark plus and cylinders are parts of an engine”

Meronymy/Holonymy (Example)

Meronymy/Holonymy

Inheritance:

A finger is part of a hand

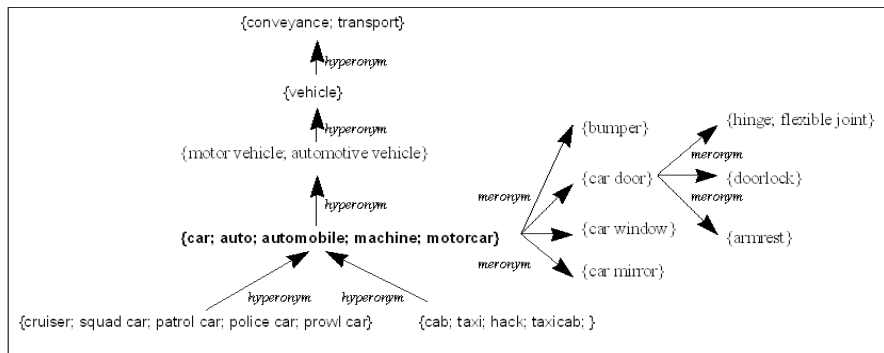
A hand is part of an arm

An arm is part of a body

=>a finger is part of a body

WordNet Hierarchy

Structure of WordNet (Nouns)



Princeton Wordnet - Car (example)

WordNet Search - 3.1

- [WordNet home page](#) - [Glossary](#) - [Help](#)

Word to search for:

Display Options:

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

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

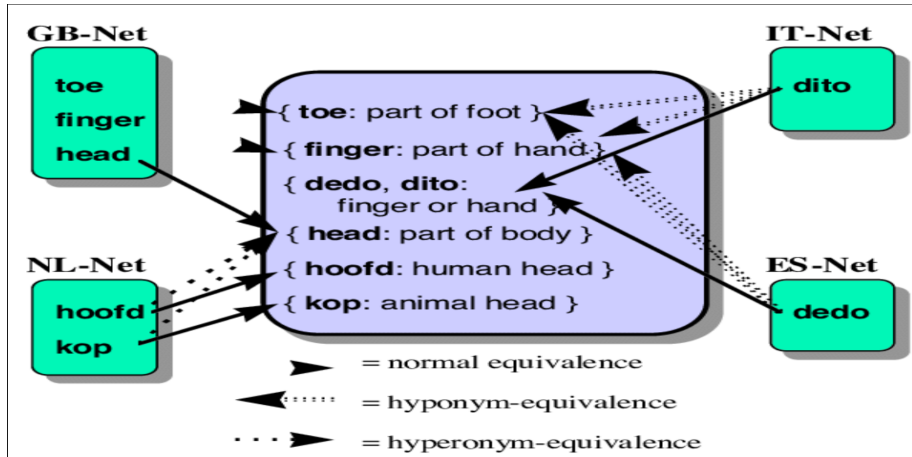
Noun

- **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"*

EuroWordNet (EWN)

- EuroWordNet is a system of lexical resources for encoding language diversity within European languages, based on WordNet.
- Each language develops its own wordnet hierarchy but they are interconnected with each other via interlingual links.
- The original EuroWordNet project dealt with Dutch, Italian, Spanish, German, French, Czech, and Estonian.
- The EuroWordNet project was completed in the summer of 1999 but further research and development as part of the EuroWordNet project is frozen.

EWN Example



LiveLanguage (LL)

- The LiveLanguage (LL) catalog provides open-access reusable lexical resources and datasets in relation to language diversity, with a particular focus on cross-lingual lexical semantics.
- The reusable datasets, along with their metadata, were produced by multiple projects involving partners from around the world across continents.
- Developed by the KnowDive research group at the University of Trento, Italy.
- Can be accessed via: LL Catalog

Live Language

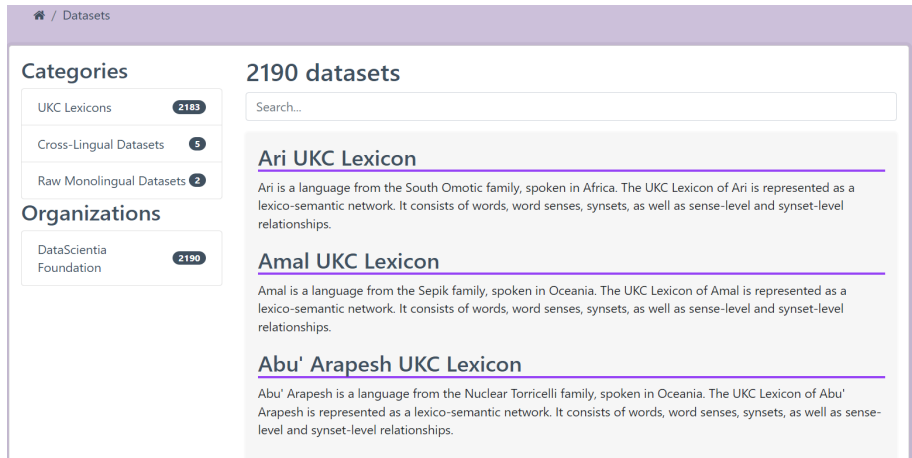
[About Us](#) [Datasets](#) [Organizations](#) [Services](#) [FAQ](#)

Explore the LiveLanguage Catalog

The UKC LiveLanguage Catalog provides open-access datasets in relation to linguistic diversity, with a particular focus on cross-lingual lexical semantics. The datasets were produced by multiple projects involving partners from around the world. Most of the data made available here can be explored and visualised on the Universal Language Core website - <http://ukc.datascientia.eu/>

[Browse All](#)

Live Language (Contd.)



Home / Datasets

Categories

- UKC Lexicons **2183**
- Cross-Lingual Datasets **5**
- Raw Monolingual Datasets **2**

Organizations

- DataScientia Foundation **2190**

2190 datasets

Ari UKC Lexicon

Ari is a language from the South Omotic family, spoken in Africa. The UKC Lexicon of Ari is represented as a lexico-semantic network. It consists of words, word senses, synsets, as well as sense-level and synset-level relationships.

Amal UKC Lexicon

Amal is a language from the Sepik family, spoken in Oceania. The UKC Lexicon of Amal is represented as a lexico-semantic network. It consists of words, word senses, synsets, as well as sense-level and synset-level relationships.

Abu' Arapesh UKC Lexicon

Abu' Arapesh is a language from the Nuclear Torricelli family, spoken in Oceania. The UKC Lexicon of Abu' Arapesh is represented as a lexico-semantic network. It consists of words, word senses, synsets, as well as sense-level and synset-level relationships.

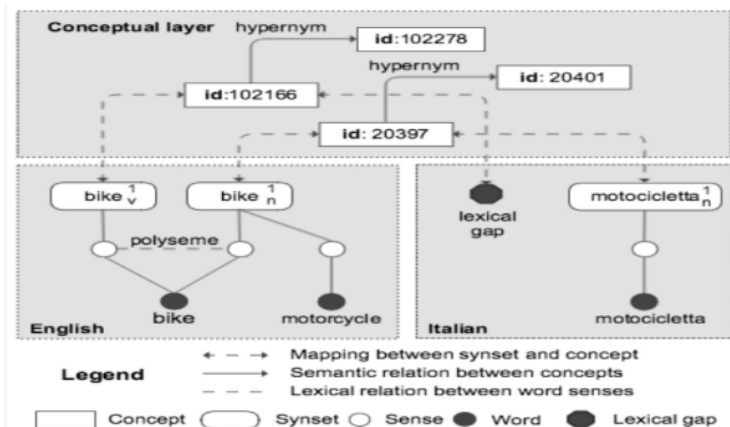
The Universal Knowledge Core (UKC)

- 1 Language Diversity
- 2 Lexical Resources
- 3 The Universal Knowledge Core (UKC)**
- 4 Language Teleontology
- 5 iTelos Language Definition Phase

The Universal Knowledge Core (UKC)

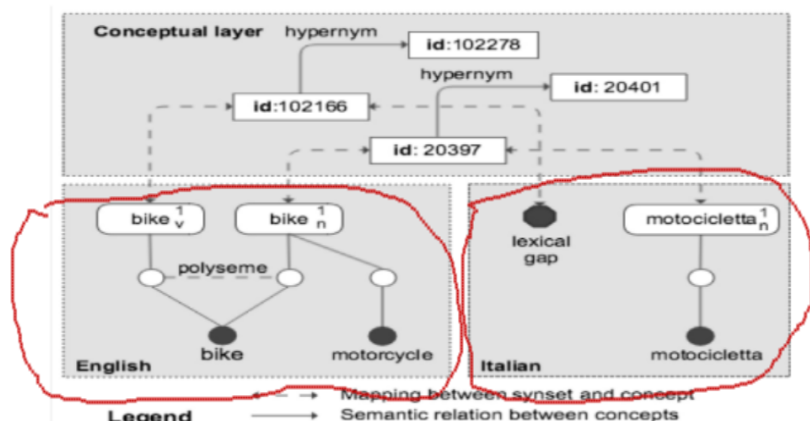
- The Universal Knowledge Core (UKC) is an expandable multilingual lexical resource encoding language diversity across 2000 natural languages and several domain languages.
- It is the reference ontology for the KGE projects allowing us to name what ontologically exists in reality.
- The UKC architecture is composed of two components:
 - Language Core (LC)
 - Concept Core (CC)
- Developed by the KnowDive research group at the University of Trento, Italy.
- Can be explored via: [UKC Website](#)

Reference Ontology - The UKC - Illustrative Example



UKC LC - Illustration

Reference Ontology - The UKC - Illustrative Example

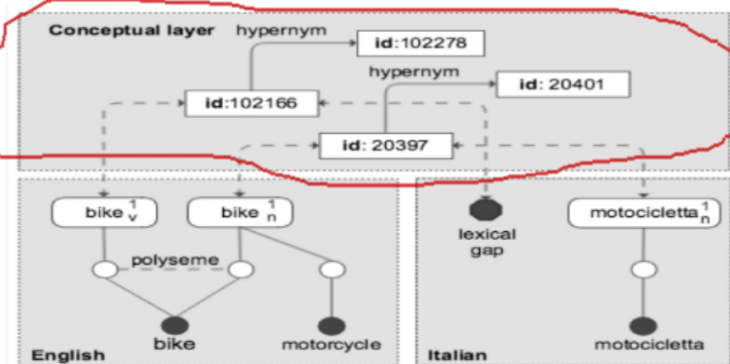


UKC LC

- We talk of the Language Core (LC), meaning the component that, in the UKC, corresponds to the PWN.
- It encodes the set of words, senses, synsets, glosses and examples supported by natural and domain languages in the UKC.
- Similarly to the PWN, in the LC each synset is univocally associated with one language and, within that language, with at least one word.
- Differently from the PWN, synsets are linked to language-independent concepts, and there is the constraint that each synset is linked to one and only one concept.
- For a concept to be created, there must be at least one language where it is lexicalized.

UKC CC - Illustration

Reference Ontology - The UKC - Illustrative Example



Legend

- ← - - - → Mapping between synset and concept
- Semantic relation between concepts

UKC CC

- The Concept Core (CC) is the UKC representation of the world and it consists of a semantic network where the nodes are language independent concepts.
- Each concept is characterized by a unique identifier (UKC GID) which distinguishes it from any other concept.
- The CC provides a uniform view over languages, it allows to compare them, to study their diversity and similarities.
- the CC makes the UKC not biased by any language and culture and, therefore, inherently open and easily extensible
- In particular, lexical gaps, namely missing concepts lexicalized in a new language can be dealt with by adding a new concept in the CC.

Language Teleontology

- 1 Language Diversity
- 2 Lexical Resources
- 3 The Universal Knowledge Core (UKC)
- 4 Language Teleontology
- 5 iTelos Language Definition Phase

Teleontology

- The meaning of the word teleontology builds on the Greek words: telos (meaning: "end, purpose"), ont (meaning: "being") and logia (meaning: "a branch of learning").
- Therefore, given what exists, the notion of a Teleontology refers to a formal, explicit hierarchical specification of objects, entities and properties representing a shared conceptualization of a specific purpose.
- A teleontology is written with a specific purpose and there is no claim of generality beyond the purpose for which it is modelled.

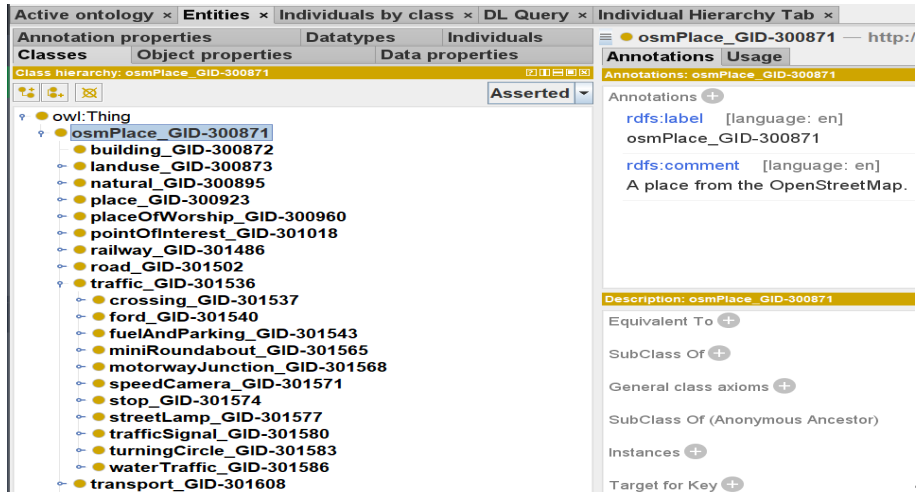
Language Teleontology (LT)

- A Language Teleontology is defined as a teleontology encoding the domain language representation of a specific KGE purpose.
- This enables representing language diversity within a specific domain and for a specific purpose.
- It represents uniquely identifiable concept names relevant to model entities and properties of a specific purpose within the scope of a knowledge graph engineering project.
- Language Teleontologies use three primary representational constructs:
 - class names, words which would be employed to later model the classes required for a KGE purpose.
 - object property names, words which would be employed to later model the object properties required for a KGE purpose.
 - data property names, words which would be employed to later model the data properties required for a KGE purpose.

Language Teleontology (Contd.)

- Three key observations:
 - First, a language teleontology should be compliant to the hierarchy of the UKC.
 - Second, a language teleontology models only concept names and not definitional assertions on such concepts, e.g.,
 - a language teleontology doesn't assert the domain and range for a particular object property name, or,
 - the domain and data type for a particular data property name.
 - Third, in practice, a LT is implemented as an ontology data structure written in OWL (web ontology language) and modelled using the protégé ontology editing software.
- In the following slides, we provide illustrative examples of the OpenStreetMap (OSM) language teleontology w.r.t class names and data property names.

OSM LT - Class (Partial) View



The screenshot shows the Protege interface with the following components:

- Active ontology:** Entities × Individuals by class × DL Query × Individual Hierarchy Tab ×
- Annotation properties:** Classes, Object properties
- Datatypes:** Datatypes
- Individuals:** Individuals, Data properties
- Class hierarchy:** osmPlace_GID-300871
- Annotations Usage:** Annotations: osmPlace_GID-300871
- Annotations:**
 - rdfs:label** [language: en] osmPlace_GID-300871
 - rdfs:comment** [language: en] A place from the OpenStreetMap.
- Description:** osmPlace_GID-300871
 - Equivalent To +
 - SubClass Of +
 - General class axioms +
 - SubClass Of (Anonymous Ancestor)
 - Instances +
 - Target for Key +
- Class Hierarchy (Left Panel):**
 - owl:Thing
 - osmPlace_GID-300871
 - building_GID-300872
 - landuse_GID-300873
 - natural_GID-300895
 - place_GID-300923
 - placeOfWorship_GID-300960
 - pointOfInterest_GID-301018
 - railway_GID-301486
 - road_GID-301502
 - traffic_GID-301536
 - crossing_GID-301537
 - ford_GID-301540
 - fuelAndParking_GID-301543
 - miniRoundabout_GID-301565
 - motorwayJunction_GID-301568
 - speedCamera_GID-301571
 - stop_GID-301574
 - streetLamp_GID-301577
 - trafficSignal_GID-301580
 - turningCircle_GID-301583
 - waterTraffic_GID-301586
 - transport_GID-301608

OSM LT - Data Property (Partial) View

Active ontology x Entities x Individuals by class x DL Query x Individual Hierarchy Tab x

Annotation properties Datatypes Individuals

Classes Object properties Data properties

Data property hierarchy: bridge_GID-301662

Annotations Usage

Annotations: bridge_GID-301662

Annotations +

rdfs:label [language: en]
bridge_GID-301662

rdfs:comment [language: en]
Is this road on a bridge? ("T" = true, "F" = false)

Characteristics: bridge_GID-301662 Description: bridge_GID-301662

Functional

Equivalent To +

SubProperty Of +
owl:topDataProperty

Domains (Intersection) +

Ranges +

owl:topDataProperty

- bridge_GID-301662
- code_GID-301663
- fclass_GID-301664
- gid_GID-301665
- latitude_GID-301666
- layer_GID-301667
- longitude_GID-301668
- maxspeed_GID-301669
- name_GID-301670
- oneway_GID-301671
- osmld_GID-301672
- population_GID-301673
- ref_GID-301674
- tunnel_GID-301675
- type_GID-301676
- width_GID-301677