KGE - Knowledge Graph Engineering

iTelos
Information models

Fausto Giunchiglia

Contents

Introduction

- iTelos aims to build and compose information models in order to create KGs.

- The models are portions of information built from a specific point of view (the Purpose). They represent specific components of the information we want to keep in the final KG.

- How to represent the information ?

- Which are the different information components considered by iTelos to build models ?

Contents

Information representation

Motivation

To represent the *world*, we use:

- Words
- Sounds
- Pictures
- .....

Representation = Concepts + Language + Knowledge + Data

For example:

Starry Night is an oil on canvas painting by Vincent van Gogh.

Contents

What are concepts?

Concepts are *alinguistic units* of human *perception* and understanding generalized from particular instances of substances which we observe.

They allow us to give an account of phenomena such as *knowledge acquisition and representation*, language understanding, inference, and categorization.

Examples:- Car, my car with nameplate FP372MK.

Contents

What is language?

Language is a *tool for representing* what is the case in the world.

We take language as a *set of terms*, each term being associated a meaning, standing for what is the case in the world.

Examples:- English, Italian, Bengali, Hindi

Contents

What is knowledge?

Knowledge is a *tool for stating* what is the case in the world.

We take Knowledge as a set of *entity types (etypes)*, each associated with a set of *properties* (Object Properties and Data Properties).

Examples:-

In the automobile domain,

etype = {Car, Motorcycle}; properties = {Body Style, Model Date, VIN}

Contents

What is data?

Data is knowledge about *individuals*.

We take data as *a set of entities*, each of a given etype, each associated with a set of *property values*.

Examples:-

My car with *VIN: FP372MK* having body style *Coupé* and model date *2020-11-25*

Contents

The problem of representation diversity

Stratification of Diversity

We take *Representation Diversity* to mean semantic heterogeneity, as organized in the *four layers*:

Conceptual Diversity (L1)

Language Diversity (L2)

Knowledge Diversity (L4)

Data Diversity (L5)

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

Summary

- We saw the different components which make up any representation

- We introduced the precise role of concepts, language, knowledge and data in terms of knowledge representation

- We saw how the above four components generate different representations, even of the same intended meaning (but not only), in different datasets, thus generating the problems trying to
  REUSE, BUILT (create & compose) and SHARE resources.

Contents

What is representation diversity?

We have Semantic Heterogeneity (e.g., in language, KBs, DBs) when there are differences in how the same real world phenomenon is represented.

Semantic heterogeneity arises whenever we have KBs and DBs developed by independent parties (in space and time).

We take Representation Diversity to mean semantic heterogeneity, as organized in the four components of concept, language, knowledge and data.

Levels of Representation Diversity

Representation diversity occurs in

1. the different concepts used to denote the same entity;
2. the different terms and meanings used in language;
3. the different entity types and the properties used;
4. the different entities and the property values used.

We categorize representation diversity in 4 levels:

- Conceptual Diversity
- Language Diversity
- Knowledge Diversity
- Data Diversity

Representation diversity is unavoidable, at all four levels.

Motivating Example

Stratification of Diversity

We take *Representation Diversity* to mean semantic heterogeneity, as organized in the *four layers*:

Conceptual Diversity (L1)

Language Diversity (L2)

Knowledge Diversity (L4)

Data Diversity (L5)

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

Contents

Conceptual Diversity (L1)

- The notion of concept is well known in *Philosophy of Mind* and in *Computational Linguistics*.

- We follow our own work and take concepts to be *unique alinguistic identifiers*.

- Concepts are organized in multiple hierarchies, in terms of '*hypernym-hyponym*' links. Eg:- Car or Vehicle ?

Language Diversity (L2)

- Languages, taken here in a very broad sense to include, e.g., *natural languages, namespaces and formal languages*.

- Linguistic phenomena like *polysemy and synonymy* allow for diverse representations of entities

- *Many-to-many mapping between words and concepts*, both within the same language and across languages

  Eg:- vso:VIN, Nameplate, Targa

11

Knowledge Diversity (L4)

- We model knowledge as a set of *entity types*, also called etypes, meaning by this, classes of entities with associated properties.

- Knowledge diversity arises from the *many-to-many mapping between etypes and the properties* employed to describe them, and can appear in two different forms.

  Eg:- Same etype Car, but associated with associated with different groups of properties.

Data Diversity (L5)

- We model data as entities each associated with property values, where properties are inherited from the etype of the entity.

- Exists because of the fact that *the mapping between entities and the property values used to describe them is many-to-many.*

  Eg:- Same entity car 'FP372MK' but with different velocities

Representation Diversity Architecture
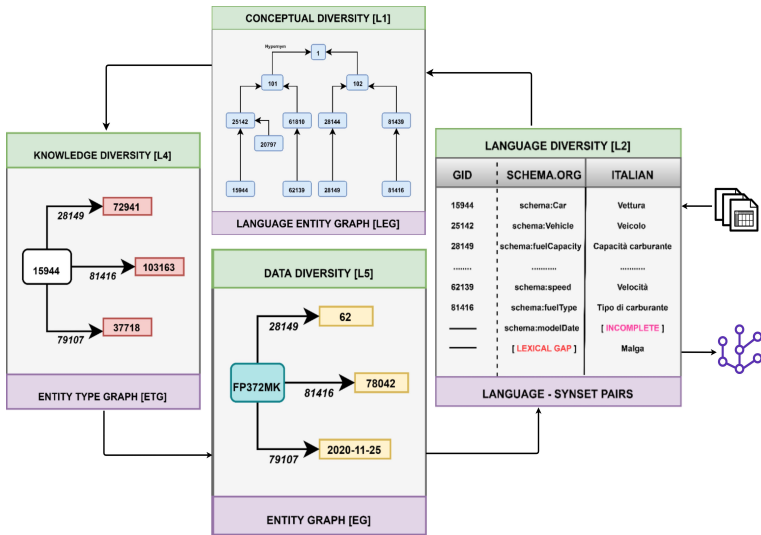
Representation Diversity Architecture (Contd.)

- The language representation layer (L2) appears first and last in the architecture. L2 enforces the input and the output dependence of the representation of data on the user language. In fact, language is the key enabler of the bidirectional interaction between users and the platform

- In the first phase, the L2 input language is translated into the system internal L1 conceptual language and the input language is only resumed during the last step, when the results of the data integration steps are presented back to the user.

- In this process, L2 is key in keeping completely distinct the multilingual user-defined data representation and the alinguistic system-level data representation.

Representation Diversity Architecture (Contd.)

- The management of conceptual diversity (L1) involves the organization of the L1 alinguistic concepts, as identified in the first step, into a *Language Entity Graph (LEG)* which codifies the semantic relations across concepts (and, therefore, among, the corresponding L2 input words).

- In order to achieve this goal we exploit, as a-priori knowledge, a multilingual lexico-semantic resource, called *Universal Knowledge Core (UKC)*

- The alignment of meanings across languages and namespaces absorbs a major source of heterogeneity present in the (Semantic) Web

Representation Diversity Architecture (Contd.)

- The net result of this phase is an LEG with the following properties:

  - the concepts identified during the first phase are all and only the nodes in this graph;

  - these nodes are annotated with the input L2 terms, across languages;

  - these nodes are organized into a hierarchy which preserves the ordering, across the links of the UKC (in the case of nouns, the synonym/ hyponym/ hypernym relations).

Representation Diversity Architecture (Contd.)

- Managing *knowledge diversity* (L4), involves the construction of a (alinguistic) *Entity Type Graph (ETG)* encoded using *only* concepts occurring in the LEG constructed during the previous two phases

- In this phase, the first step is to distinguish concepts into *etypes* and *properties* (both object properties and datatype properties) while the second step is to organize them into a *subsumption hierarchy*

- In the fourth representation layer (L5), we tackle data diversity via an *Entity Graph (EG)*, namely, a *data-level knowledge graph*, by populating the *ETG* with the entities extracted from the input datasets.

Representation Diversity Architecture (Contd.)

- The EG is constituted of a backbone of L1 alinguistic ids, each annotated with the input L2 terms where, for each L2 term, the system remembers the dataset it comes from.

- This mechanism is implemented via a *provenance* mechanism which applies to all the input dataset elements, both at the schema and at the data level.

- One major advantage of our approach is that the combinatorial explosion deriving from the interaction of the four different types of diversity is avoided and *the complexity of the data integration problem reduces to the sum of the complexity of each layer.*

Did we forget L3 ?

- **No, L3 exists !**

- L3 is a particular kind of knowledge resource, called *Teleology*.

- Teleologies uses L2 elements within an extended formalization of entity relationship (ER) models, with the objective of modeling a purpose specific representation of such concepts as well as the exiting relations among them.

- A teleology is the **contextual model for the Purpose** considered.

- iTelos composes the Purpose's teleology with already existing, well known, ontologies in order to create the unique knowledge structure for the final KG.

A dedicated lecture will define teleologies and their usage in detail.

In the next lectures ...

- **Models implementation**

- How the different kinds of models are concretely shaped in iTelos ?

- In which form they appear along the methodology ?
- How iTelos is able to create and compose models ?

Fausto Giunchiglia

iTelos
Information models