KDI Knowledge and Data Integration

Diversity in Stratified Representation

Fausto Giunchiglia, Mayukh Bagchi

Contents

1 **Representation Diversity**

2 Stratified Knowledge

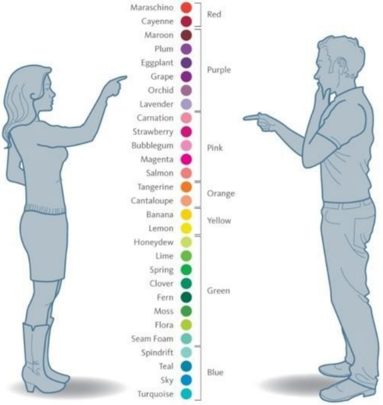
3 Summary

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

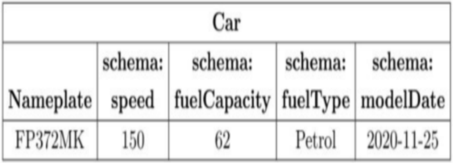
We take *Representation Diversity* to mean semantic heterogeneity, as organized in the *four layers*:

Conceptual Diversity (L1)

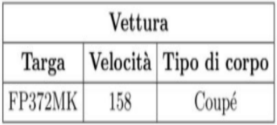
Language Diversity (L2)

Knowledge Diversity (L4)

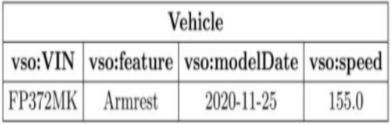
Data Diversity (L5)



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

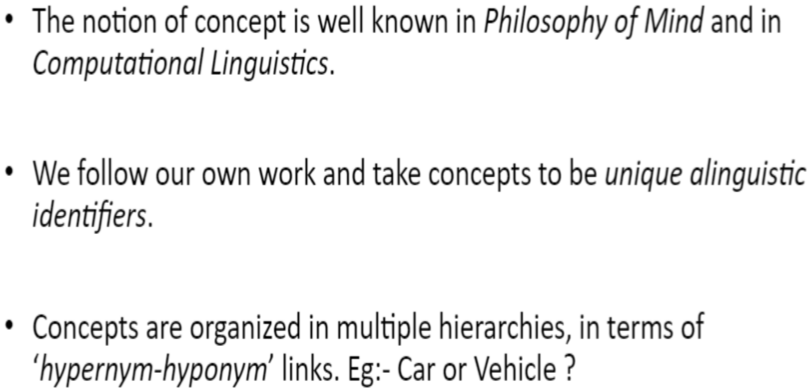
Contents

1 Representation Diversity

2 Stratified Knowledge

3 Summary

Conceptual Diversity (L1)

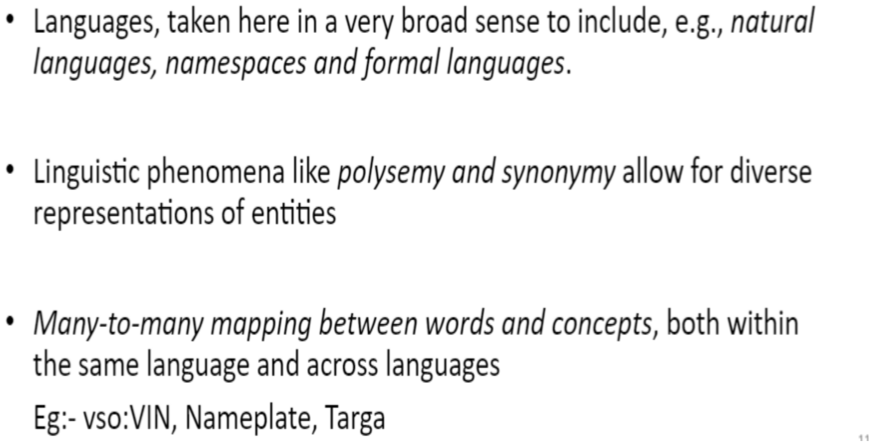


- The notion of concept is well known in *Philosophy of Mind* and in *Computational Linguistics*.

- 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*.

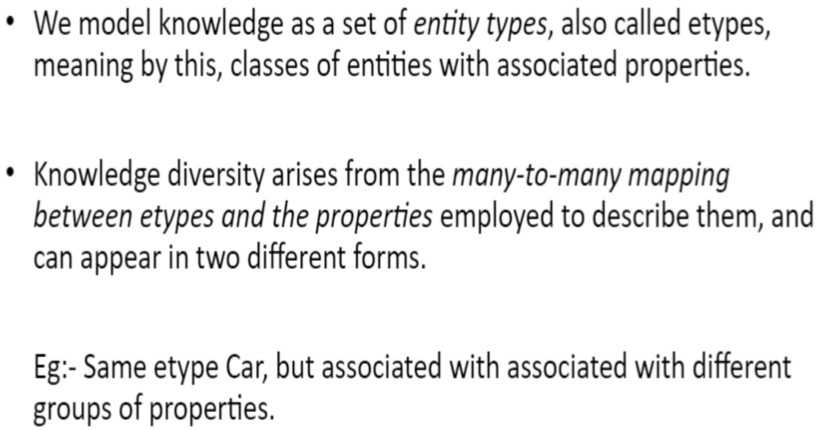
- Linguistic phenomena like *polysemy and synonymy* allow for diverse representations of entities

- *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)

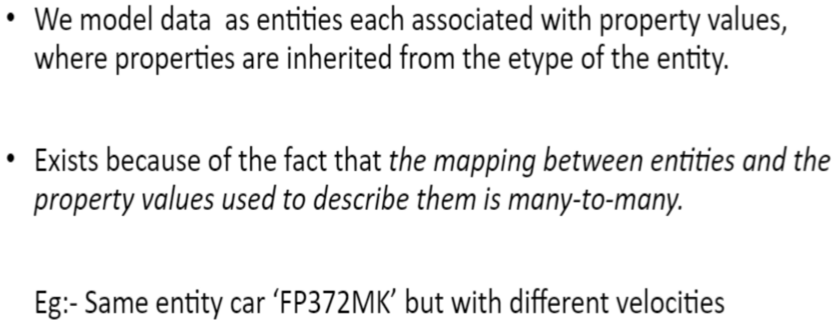


- We model knowledge as a set of *entity types*, also called etypes, meaning by this, classes of entities with associated properties.

- 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.

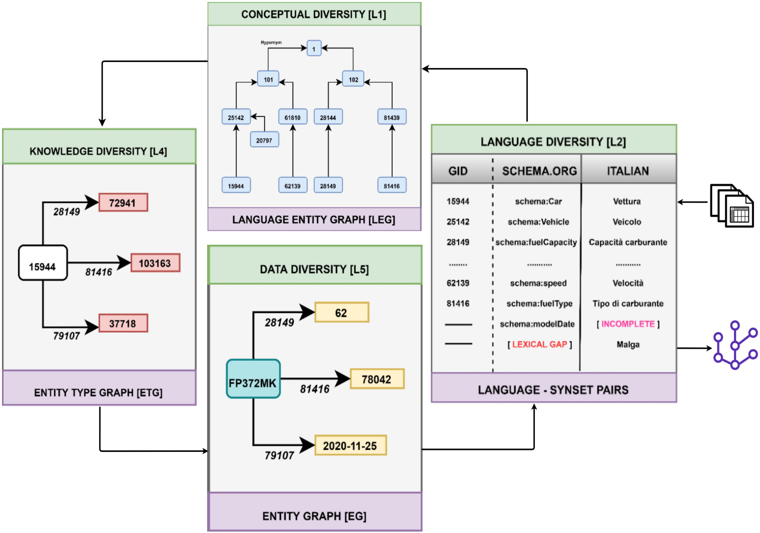
Data Diversity (L5)



- 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



GID	SCHEMA.ORG	ITALIAN
15944	schema:Car	Vettura
25142	schema:Vehicle	Veicolo
28149	schema:fuelCapacity	Capacità carburante
....	....	....
62139	schema:speed	Velocità
81416	schema:fuelType	Tipo di carburante
——	schema:modelDate	[INCOMPLETE]
——	[LEXICAL GAP]	Malga

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.*

Contents

1 Representation Diversity

2 Stratified Knowledge

3 **Summary**

Summary

- We learnt about the different genres of representation diversity impeding a data integration task via the motivating example from automotive domain

- We saw in detail how our stratified data integration approach reduces the complexity of the data integration problem to the sum of the complexity of each representation layer.

KDI Knowledge and Data Integration

- Fausto Giunchiglia, Mayukh Bagchi
- Diversity in Stratified Representation