On Knowledge Diversity

Fausto GIUNCHIGLIA [a] and Mattia FUMAGALLI [a]
[a] *Department of Information Engineering and Computer Science (DISI)*
University of Trento, Italy

Abstract. In this short paper we introduce two main elements of a general methodology, called *iTelos*, for the management of *knowledge diversity*. The first is *Knowledge Lotuses*, a general purpose tool for the *representation* of knowledge diversity, while the second is a *set of metrics* which allow to *quantify* it, as it occurs within and across knowledge resources.

Keywords. knowledge diversity, knowledge representation,, context

1. Introduction

We usually talk of *Semantic Heterogeneity* meaning the phenomenon which arises when multiple knowledge resources, e.g., ontologies or schemas, for the same domain, present differences in how the intended meaning is represented, most often as a consequence of the fact that they have been developed independently. Managing this phenomenon is crucial in order to enable the *Semantic Interoperability* of knowledge resources. The key intuition underlying most previous work is to reduce the input representations to a given reference representation, e.g., an ontology, still preserving the intended meaning. This problem has been extensively studied in the literature, leading to a substantial amount of results. As an example, LOV, LOV4IoT,[1] three among the most relevant repositories of reference knowledge resources, collectively contain around 800 such resources, some of which contain thousands of elements.

This work has gone a long way with many success stories, in particular in high value, highly formalized, domains, e.g., health, manufacturing. However, a general solution to the semantic interoperability problem, applicable with sustainable costs and time effort, is yet to be found. The difficulties which arise are multifaceted. Some are related to the fact that different resources only consider different *partial* aspects of a domain, or that they represent it at different levels of *abstraction* and/or *approximation*. Furthermore, last but not least, no matter how it is built, any resource will be hardly *reusable* in novel contexts and it will most often need *to be adapted* and *evolve* in time.

These difficulties are deeply rooted into the nature of knowledge. People adapt their representations of the world as a function of their goals, focus and many other factors [1]. These local representations, though useful, are the key cause of semantic heterogeneity, this phenomenon being in fact unavoidable. It is simply impossible to construct a *finite representation* capable of capturing the *infinite richness of the world* and also the *infinite ways*, provided by language, *to describe finitely* (some aspect of) *the world itself*. Thus, on one hand, for any chosen representation there will always be some aspect of the world

[1] `https://lov.linkeddata.es/dataset/lov`, `http://lov4iot.appspot.com/`.

which is not captured and, on the other hand, there will always be an alternative way to represent the same aspect of the world. This phenomenon is further complicated by the fact that *world diversity* and *representation diversity* are independent, the first being rooted in the world itself, and the second in how people think about it.

In this paper, we propose a novel methodology, called *iTelos*,[2] for the management of the *diversity of knowledge* [2], as it arises from the combination of the world and representation diversity, independently of whether it comes from humans or from machines. (Notice how semantic heterogeneity is just one form of knowledge diversity). *iTelos* is based on the following three key intuitions: *(i)* it should support all phases of the *knowledge life cycle*, end-to-end. Notice how iTelos considers, among others, both the generation of a resource from scratch and its generation from existing resource (the latter being the focus of the semantic interoperability problem); *(ii)* It should make explicit the resources' representation choices as well as their motivations. The current version of *iTelos* considers two such motivations, namely, a set of generalized questions [3] and, the set of resources that are reused [4]; *(iii)* it is crucially based, at the knowledge level, on the notion of *teleology* [5], where teleologies are knowledge resources constructed similarly to ontologies, but by *making explicit* their underlying representation choices. The key idea is to use this information as the basis for the (semi-)automatic low-effort reuse of ontologies.

Within this framework, our goal below is to introduce two new key elements of *iTelos*, namely *Knowledge Lotuses*, as a general tool for representing and visualizing knowledge diversity (Section 2) and an initial *set of metrics* which *quantify* the level of diversity within and across teleologies (Section 3).

2. Representing Diversity

Let us assume that a teleology represents knowledge in terms of (*types of*) *entities* (e.g., *Person*, *Place*, *Event*), each being associated with a set of *properties* (e.g., *birth-date*, *height*, *near-to*, *father-of*, *has-capital*), as it is the case in, e.g., knowledge graphs or relational models. As from Formal Concept Analysis (FCA) [6], we formalize teleologies as contexts, where we define a context C as $C = \langle E_C, P_C, I_C \rangle$, with $E_C = \{e_1, ..., e_n\}$ being the set of entities, $P_C = \{p_1, ..., p_n\}$ being the set of properties of C, and I_C being $I_C = \{(e, p) \in I_C \mid p \text{ is a property of } e\}$. I_C is a *Galois connection* [6]. The set of properties associated to an entity is called its *intention*, while we talk of an entity e being in the domain of a property p, formally *dom(p)*. Thus, for instance, the entity "Person" can be in the domain of the properties "address" or "name", while the property "address" may occur with the entities "Person", or "Building". Following the FCA notation, Table 1 reports a set of entities (left) with corresponding properties (top) from Schema.org.rdf Version 3.5. The value boxes with crosses represent I_C.

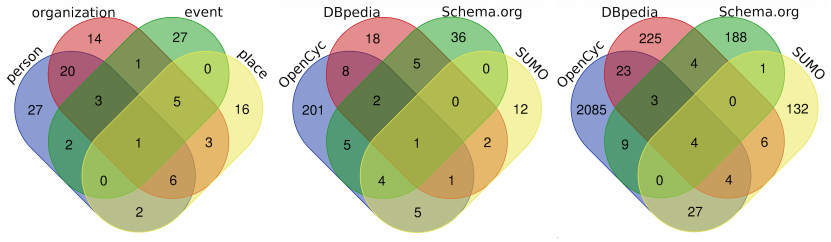
We represent the diversity which occurs within and across teleologies with *Knowledge Lotuses*. Fig 1 provides three knowledge lotuses for (parts of) four state-of-the-art knowledge schemas, namely OpenCyc, (the) DBpedia (ontology), Schema.org, and

[2] From the Greek word *telos*, meaning "end, purpose". The "*i*" stands for "integration".

Table 1. A context for (a portion of) Schema.org

schema.org (representation) context		Properties						
		actor	*actors*	*add*	*added*	*additional*	*address*	*caption*
entity	schema:PriceSpecification				×			
	schema:Offer			×				
	schema:GeoCoordinates						×	
	schema:MusicGroup							
	schema:Person					×	×	
	schema:VideoObject	×	×					×
	schema:OrganizationRole							
	schema:Role				×			

SUMO.[3,4] Knowledge lotuses are *Venn Diagrams*[5] defined as follows. Let us assume that we are interested in analyzing the diversity of certain a set of contexts C, e.g., the four resources mentioned above. Knowledge lotuses model the three core elements of contexts, namely, *(i)* the set of contexts themselves and, for each context, *(ii)* its set of entities and, for each entity, *(iii)* its set of properties. The key intuition is to fix one of these three elements (namely the context(s), or the entity(ies), or the property(ies)) and then, under this assumption, to study the diversity of the second element against the third element. Clearly, several combinations are possible and each of them provides a different perspective on diversity. As an example, let us consider the three cases in Fig 1.



(a) Schema.org (b) Organization (c) All properties

Figure 1. (a) Schema.org shared properties across four entities; (b) "Organization" shared properties across four resources; (c) shared entities across four resources (independently of their properties).

Lotus (a) fixes the context (Schema.org) and it represents the diversity of entities in terms of their (un)shared properties. The dual case of comparing properties in terms of the entities in their domain is also possible. These types of lotuses represent the *diversity internal to a context, in terms of their entities or their properties.* Lotus (b) fixes the entity (Organization) and it represents the diversity of teleologies in terms of their (un)shared properties (for that entity). The dual case of comparing properties in terms of the teleologies where they occur is also possible. These types of lotuses represent the *diversity across teleologies, for any given entity*. Lotus (c) fixes the properties (all of them) and it represents the diversity of teleologies in terms of the (un)shared entities. The dual case of comparing entities in terms of the teleologies where they occur is also possible. These types of lotuses represent the *diversity across teleologies, for any given property*.

For instance, looking at Fig1(a), in Schema.org, "Person" and "Organization" share 30 property terms, while they are distinguished by 31 and 23 terms respectively. Looking

[3] `www.cyc.com`, `wiki.dbpedia.org`, `www.schema.org`, `www.adampease.org`.

[4] The data in Fig 1, as well as the quantitative analysis below, have been generated via a simple NLP pipeline which performs the following main steps: a) split a string every time a capital letter is encountered (e.g., birthDate → birth and date); b) lower case all characters; c) filter out stop-words (e.g., hasAuthor → author).

[5] All the lotuses in Fig 1 represent four sets. Simpler/complex lotuses can be depicted to represent the diversity of lower/ higher numbers of resources.

at Fig 1(b), the four representations of "Organization" share only one property, while two of them, i.e., OpenCyc and DBpedia share 12 properties.

One observation. Consider Fig 1(c). If one looks at the central part, there are only four entities which are shared by all resources. These entities are *Event*, *Place*, *Person* and *Organization*, namely the entities for time and space, i.e., the two *a priori* of perception [7] and the two arguably most common types of *agenthood*. Dually, if one looks at the fringes, most entities are defined in only one resource (e.g., 2085 in OpenCyc) this being motivated by the different focus. Thus, for instance, Schema.org is more focused on information objects, while DBpedia contains information about biological species. Despite the fact that these four resources are arguably general purpose and that, therefore, they somehow take a similar view of the world, they present a very *low level of unity* (in the shared part) together with *a high level of diversity* (in the unshared parts). This is further evidence of the fact that there is no such notion of an observer independent representation of the world. Analogous argumentations, all providing motivations for a quantitative study of diversity, could be given also for the other lotuses and for other knowledge resources.

3. Quantifying Diversity

We analyze first the diversity *within a resource* and then *across multiple resources*. Our key insight for analyzing a resource internal diversity is based upon Rosch's *cue validity* [8]. This notion was used to define the set of *basic level categories*, namely those categories which maximize the number of characteristics shared by their members and minimize the number of characteristics shared with the members of their sibling categories. Following Rosch, we define *the cue validity of a property p w.r.t to an entity e*, also called cue_p-$validity$, as:

$$Cue_p(p,e) = \frac{PoE(p,e)}{|dom(p)|} = c \in [0,1] \tag{1}$$

with $|\ X\ |$ being the cardinality of the set X and *PoE(p, e)* being defined as:

$$PoE(p,e) = \begin{cases} 1, if\ e \in dom(p) \\ 0, otherwise \end{cases} \tag{2}$$

(p, e) returns 0 if p is not associated with e and *1/n*, where n is the number of entities in the domain of p, otherwise. In particular, if p is associated to only one entity its cue_p-$validity$ is maximum and equal to one. Given the notion of cue_p-$validity$ we define the notion of *cue validity of an entity*, also called cue_e-$validity$, as the sum of the cue validities of the properties associated with the entity, namely:

$$Cue_e(e) = \sum_{i=1}^{|prop(e)|} Cue_p(p_i,e) = c \in [0, prop(e)] \tag{3}$$

where *prop(e)* is the set of properties which are associated with e. The intuition is the same as Rosch's: the entities with higher cue_e-$validity$ will be the easiest to recognize. In fact, the cue_e-$validity$ increases with the number of the properties, while decreasing, for each property, with the number of entities which share that property.

However, the cue_e-$validity$ does not tell us anything about how many properties an entity shares with other entities in the same context. To make an example, assume that

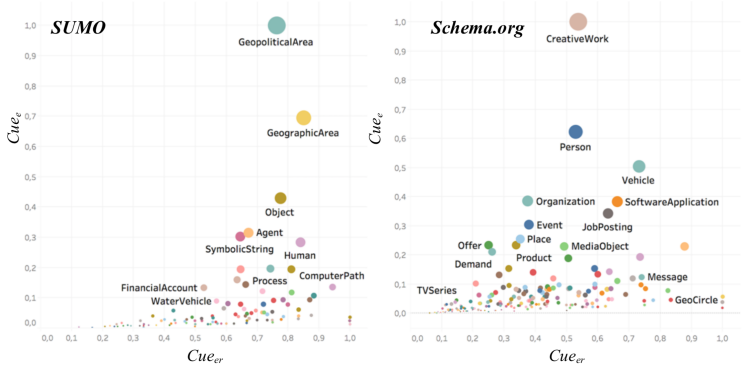


Figure 2. Cue_e and Cue_{er} of SUMO and Schema.org, where the size of the circles is proportional to the number of properties for that entity.

we have two entities and two properties. Assume the following two situations: *(i)* both entities share both properties and *(ii)* the two entities are each associated to one property. In both cases the cue_e-$validity$ of the two entities is 1 but, while in the first case they are indistinguishable, in the second case they are highly identifiable. We capture this distinction via the notion of cue_{er}-$validity$, as:

$$Cue_{er}(e) = \frac{Cue_e(e)}{|prop(e)|} = c \in [0,1] \tag{4}$$

The higher the cue_{er}-$validity$ the more distinguishable an entity is, for a given value of the cue_e-$validity$. Thus, for instance, in the first case in the example above, the cue_{er}-$validity$ of the both entities will be 0.5 while in the second case it will be one. As an example let us analyze the internal diversity in SUMO and Schema.org as from Fig 2, where the x-axis and y-axis are cue_{er} and cue_e, respectively. A few observations are in order. The first is that both resources have a few outlier entities (e.g., "GeopoliticalArea", "Person" and "CreativeWork") with higher values of cue_e-$validity$. Furthermore, there seems to be a pattern by which, the more the cue_e-$validity$ decreases, the more entities there are with the same cue_{er}-$validity$, this meaning the fact that there is an area with a cloud of entities which are not easy to distinguish among one another. Furthermore, the outlier entities are the only ones which high values of cue_e-$validity$. By looking then into the specifics, it seems that the outliers are mostly those entities of higher interest (e.g., "CreativeWork" and "Person" in Schema.org) as, maybe, was to be expected (in anything we do, we all tend to focus on what is of highest interest). This confirms once more the role of diversity in capturing not only what is formalized (as well in knowledge lotuses) but also the quality of the formalization (via metrics). We define the basic diversity measure across resources via the *Jaccard index* [9]. Let C_A and C_B be two contexts, with their sets of properties, $prop(C_A)$ and $prop(C_B)$. Then, we define the *similarity of two contexts C* as follows:

$$Sim_c(C_A, C_B) = \frac{|prop(C_A) \cap prop(C_B)|}{|prop(C_A) \cup prop(C_B)|} = c \in [0,1] \tag{5}$$

$Sim_c(C_A, C_B)$ is a *symmetric* measure which tells us how much is (not) shared across two resources. If this measure is 1 then the two resources coincide, if it is 0 then they are disjoint. As an example of use, take the two contexts to be a single entity, for instance as formalized in two different resources. Clearly the value of this resource is independent of the actual name of the entity itself. This measure will thus allow, for instance, to realize

that the two input entities are the same despite the fact that they have different names and also the vice versa. Fig 3 below shows some examples of similarity of entities from SUMO and Schema.org. In the y-axis we have the value of $Sim_C(C_A, C_B)$ with C_A being the entity written above the graph as formalized as formalized in either SUMO (SU) or Schema.org (SC), while in the x-axis we have other entities from both resources (each entity being taken as C_B), in decreasing order of similarity. For instance, "Apartment" in SC is essentially a synonym of "SingleFamilyResidence" in SC with a similar situation with "MoveAction". Notice how the two entities in SC, which are synonyms of "MoveAction", are, by transitivity, also synonyms (where it is not clear whether this was actually what the modeler really wanted to do). The other diagrams report much lower levels of similarity, where for instance "Person" in SC has the highest similarity with "Organization" as in SC or SU, namely the other entity for agenthood.

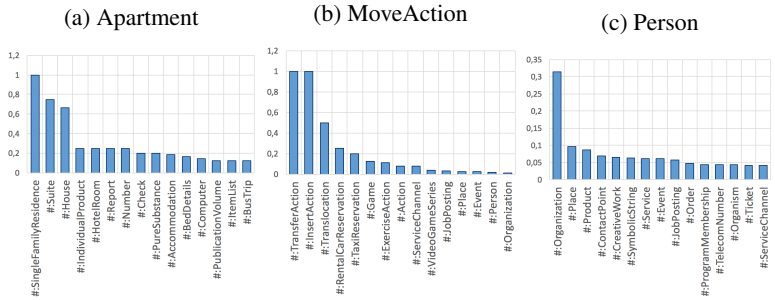


Figure 3. Similarity between different entities in Schema.org and SUMO.

4. Conclusion

We see this as just the beginnings of a quantitative study of knowledge diversity. We see potential from a *scientific point of view*, towards the study of knowledge as an emerging natural phenomenon, but also from an *engineering point of view*, towards a widespread reuse of existing resources (teleologies). This work is part of a long term effort aimed at providing *iTelos*, a general methodology for managing knowledge diversity and, in particular, for performing knowledge and data integration in a cost-effective way.

References

[1] F. Giunchiglia and M. Fumagalli. Concepts as (recognition) abilities. In *FOIS*, 2016.

[2] F. Giunchiglia. Managing diversity in knowledge. In *IEA/AIE*, page 1, 2006.

[3] U. Chatterjee, F. Giunchiglia, D. P. Madalli, and V. Maltese. Modeling recipes for online search. In *OTM On the Move to Meaningful Internet Systems*, pages 625–642. Springer, 2016.

[4] S. Das and F. Giunchiglia. Geoetypes: Harmonizing diversity in geospatial data (short paper). In *OTM On the Move to Meaningful Internet Systems*, pages 643–653. Springer, 2016.

[5] F. Giunchiglia and M. Fumagalli. Teleologies: Objects, actions and functions. In *ER 2017*, pages 520–534. Springer, 2017.

[6] B. Ganter and R. Wille. *Formal concept analysis: mathematical foundations*. Springer, 2012.

[7] Immanuel Kant. Critique of pure reason (translated and edited by p. guyer & a. w. wood). 1998.

[8] E. Rosch. Principles of categorization. *Concepts: core readings*, 189, 1999.

[9] R. Real and Juan M Vargas. The probabilistic basis of jaccard's index of similarity. *Systematic biology*, 45(3):380–385, 1996.