

# Measuring and controlling knowledge diversity

Yasser Bourahla<sup>1</sup>, Jérôme David<sup>1</sup>, Jérôme Euzenat<sup>1</sup> and Meryem Naciri<sup>1</sup>

<sup>1</sup>Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG, 38000 Grenoble, France

## Abstract

Assessing knowledge diversity may be useful for many purposes. In particular, it is necessary to measure diversity in order to understand how it arises or is preserved; it is also necessary to control it in order to measure its effects. Here we consider measuring knowledge diversity using two components: (a) a diversity measure taking advantage of (b) a knowledge difference measure. We present the general principles and various candidates for such components. We discuss how these measures may be used to generate populations of agents with controlled levels of knowledge diversity.

## Keywords

Knowledge diversity, Diversity measure, Ontology dissimilarity, Diversity control, Entropy

## 1. Introduction

Agents may hold different knowledge. This characterises the knowledge diversity of an agent population. In general, diversity is an important asset. It has been shown, in different contexts, that groups of agents with diverse abilities have better problem solving skills those with high abilities [1, 2, 3]. In an evolutionary context, diversity is considered to have influence on species resilience [4].

However, knowing that agents in a population are diverse does not tell how much diverse is the population's knowledge, nor which population has more diverse knowledge. There are various reasons to assess agent populations' knowledge diversity, in particular studies in social modelling. Our own work aims at performing experiments for which we have to *measure* knowledge diversity [5], because we want to characterise those factors that promote or inhibit diversity, and we have to *control* knowledge diversity, because we want to characterise its influence on other factors, e.g. robustness. Of course, the diversity that is measured should be the same as that which is controlled.

Hence we are in need of formal knowledge diversity models applicable to formal knowledge representations [6]. In this paper, we focus on both measuring and controlling knowledge diversity in a coherent way. Measuring knowledge diversity may be split in two components: (a) taking advantage of proved diversity measures (b) relying on knowledge-specific structures, and in particular ontology dissimilarities. We show how this can be achieved in an integrated way. We do not provide *the* single best measure, but instead an ordered set of measures and methods which balance accuracy and complexity. Controlling knowledge may be grounded on such measures.

---


*The Eighth Joint Ontology Workshops (JOWO'22), August 15-19, 2022, Jönköping University, Sweden*

✉ Yasser.Bourahla@inria.fr (Y. Bourahla); Jerome.David@univ-grenoble-alpes.fr (J. David);

Jerome.Euzenat@inria.fr (J. Euzenat); Meryem.Naciri@inria.fr (M. Naciri)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

After stating the problem to solve (§2), we discuss related work addressing it (§3). We first provide a concrete example of ontology dissimilarities based on the semantics of the representation language and one specific way to categorise ontologies (§4). We then introduce a simple way to measure knowledge diversity among agents based on such measures (§5). This view is refined by segmenting the knowledge space into a priori categories and considering the distribution of agents with respect to these (§6). Finally, we discuss simple ways to control knowledge diversity among a population of agents with respect to the proposed measures (§7).

## 2. Problem statement

The problem which is considered here is to associate a number to the knowledge diversity of a population of agents (measure) and to associate knowledge to a population of agents so that their diversity is close to such a number (control). We only consider agent populations of the same size.

In principle, a population  $A \in \mathcal{A}$  of agents is characterised by a set of features  $F$ . Each  $a \in A$  may be considered a point in the space induced by  $\times F$ . In our case, there is a single, but complex, feature  $f$  which associates to each agent its knowledge (in this paper, we assume that it is an ontology  $o \in \mathcal{O}$  expressed in a description logic [7]).

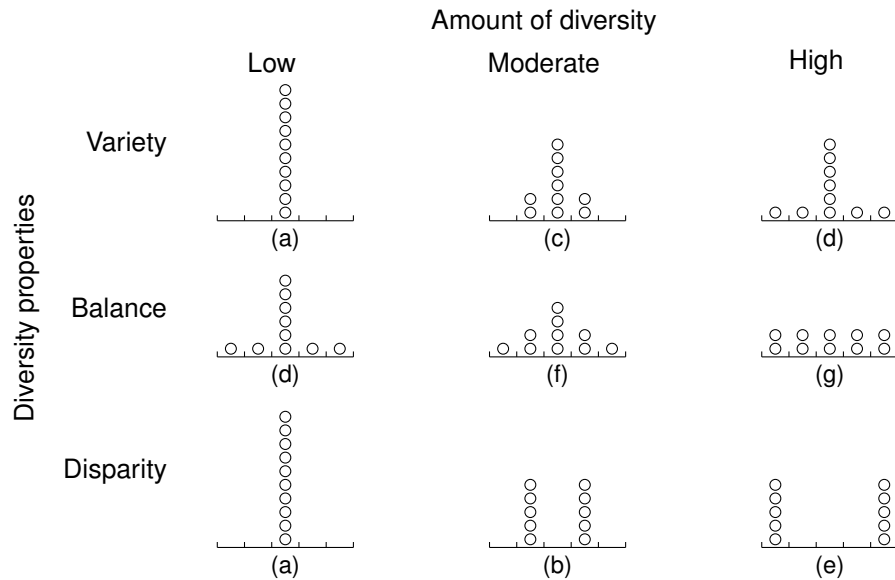
Given an agent population  $A$ ,  $O_A = \{f(a); a \in A\}$  is the multiset of their ontologies representing the distribution of these ontologies in the population. If one replaces some of these ontologies by others, then this multiset contains 0 or more occurrences of each ontology. The number of different ontologies used by the population is noted  $|O_A|$  and the number of occurrences of ontology  $o$  in  $O$  is  $\#o$  (also called the *abundance* of  $o$ ). By extension, the cardinal of a distribution  $S$  is noted  $\#S$ , e.g.  $\#O_A = \sum_{o \in \mathcal{O}} \#o = |A|$ .

Accounting for diversity may be achieved by defining an index  $\delta : \mathcal{A} \rightarrow \mathbb{R}$  of diversity within a population or a partial order  $\preceq \subseteq \mathcal{A} \times \mathcal{A}$  indicating that a population is more diverse than another. Such a diversity measure ( $\delta$ ) is an absolute measure within a population. It is also possible to consider diversity across populations. For that purpose, it is sufficient to define an order relation  $\preceq \subseteq \mathcal{A} \times \mathcal{A}$  (a population's knowledge is less diverse than another). When a diversity measure  $\delta$  has been defined, it is easy to test if  $\delta(A) \leq \delta(A')$  and then decide that  $A \preceq A'$ . It may also be possible to define such an order directly.

## 3. Related work

With respect to this problem, three lines of related works may be considered: how this is approached in biology, how this is approached in social sciences, and how knowledge difference is measured in knowledge representation.

**Measuring biological diversity.** In genetics, diversity is usually obtained by measuring a distance between observed biological objects [8, 9]: (DNA) sequences (blast), genes or proteomes. This measures the distance between individual characteristics or possibly species (represented by one representative object or *specimen*). It has to be generalised to populations. In ecology, many measures of diversity exist [10]. One important line of these is based on the probability of a random individual to be in a category and rely on entropy [10, 11].



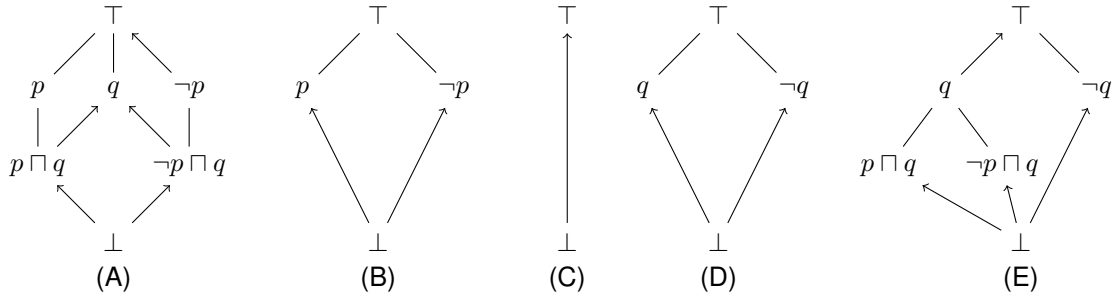
**Figure 1:** Presentation of the diversity properties of [1] inspired by Figure 1 of [14]. 10 objects are distributed within 5 different categories. Variety considers the number of instantiated categories, balance, the evenness of the distribution, and disparity, the distance between the categories (here taken as their linear position). These dimensions are independent.

**Measuring social diversity.** Measuring social and human diversity may be a more controversial issue [12]. Because the number of distinguishing features can be large, it is necessary to reduce these to identifiable categories. This may be achieved by clustering [13] or by using predefined questionnaires, localising people on a predefined space partition. [1] identifies three properties to take into account in measuring diversity (see Figure 1) which may be summarised by:

- Variety: how many categories are represented?
- Balance: how many representatives of each category are there?
- Disparity: how different are these categories?

[14] provides a different interpretation of disparity as the way an amount of resources is shared among a population depending on the feature values. Although, the latter is important in social sciences, knowledge diversity may be better measured in terms of the three former ones.

**Knowledge measures.** Concerning knowledge representation, we have the possibility to directly access agent knowledge, like DNA sequences. Hence, it should be possible to measure knowledge diversity [16]. [15] considers how to compute diversity by extracting data from graphs called ‘heterogeneous information networks’. This could be applied directly to RDF graphs. It consists of specifying random walks from which collecting the data and applying an entropy measure on the probability distribution of the collected data. Beyond data, various measures have been developed to compute distances or dissimilarities between knowledge representations. These may be used to assess disparity. In particular, ontology dissimilarities have been designed based on lexical (4.2 of [17]), syntactic [17, 18], vector-space (4.1 of [17]), alignment-based (using explicit relations between ontologies [19]), instance-based [16], semantic (based on models and



**Figure 2:** Five ontologies to be compared. Each of the concept presented in these corresponds to the named classes of the ontologies.

closure) [20]. Once measures between individual knowledge are available, classical descriptive statistic measures can be used to aggregate them at the population level.

We will build on such ontology dissimilarities to assess and control knowledge diversity in a set of agents.

## 4. Knowledge dissimilarity and categories

In order to provide precise examples, this section introduces specific ways to measure how far away knowledge representations can be and how they can be grouped into categories. These techniques have been used to measure diversity in the experiments of [5]. However, this example applies on a specific use case, in which knowledge is expressed as description logic ontologies and diversity is based on the semantics of named classes. This is not supposed to be the only answer to measuring knowledge differences. It is an example showing that knowledge can be taken seriously, through its semantics, when measuring diversity. All techniques briefly mentioned in Section 3, from the simplest to the most sophisticated, may be used instead.

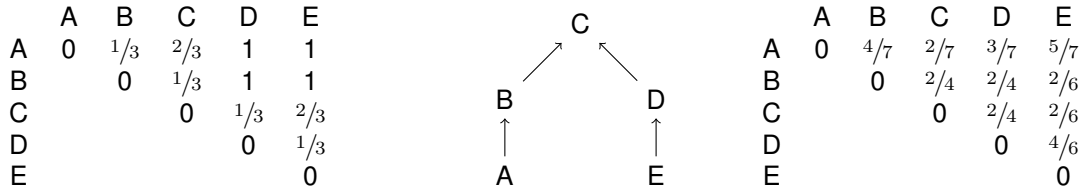
### 4.1. Ontologies

Knowledge is expressed through a very limited description logic [7] based on a finite set of properties  $P \neq \emptyset$ . For simplicity, the properties are considered Boolean, i.e. an object either has a property  $p \in P$  or it does not. The grammar of class descriptions is:

$$C := \top \mid \perp \mid \exists p.\top \mid \forall p.\perp \mid C \sqcap D \mid C \sqcup D \mid \neg C$$

Constraints on properties may be  $\exists p.\top$  (noted  $p$ : the class of objects having property  $p$ ) or  $\forall p.\perp$  (noted  $\neg p$ , the class of objects not having property  $p$ ). Named classes are defined through axioms of the form ‘name  $\equiv$  description’. The set  $C(o)$  of the named classes of ontology  $o$  contains  $\top$  (the class of all objects) and  $\perp$  (the empty class). We assume that ontologies follow the unique name assumption: two equivalent classes cannot have different names.

Figure 2 presents five simple such ontologies. We use them to illustrate dissimilarity between ontologies and its use to compute diversity among agents.



**Figure 3:** Semantic dissimilarity measures between the ontologies of Figure 2. Graph-based (left) and named-class-based (right) and the transitive reduction of the subsumption graph  $\langle O, \sqsubseteq \rangle$  (middle).

## 4.2. Semantic dissimilarity between ontologies

In [5], the dissimilarity  $d$  between two ontologies  $o$  and  $o'$  is defined as:

$$d(o, o') = 1 - \frac{|\{(c, c') \in C(o) \times C(o') \mid o \cup o' \models c \equiv c'\}|}{\max(|C(o)|, |C(o')|)}$$

i.e. the proportion of named classes in the largest ontology with no equivalent class in the other.

If two ontologies are semantically equivalent, from the standpoint of their named classes, then their dissimilarity will be 0. For instance, consider the ontology  $E$  of Figure 2 in which  $\neg p \sqcap q$  is expressed as  $\neg(p \sqcup \neg q)$ , then its dissimilarity to  $E$  will be 0.

The use of this dissimilarity on the ontologies of Figure 2 is given in Figure 3 (right).

## 4.3. Graph-based dissimilarity between ontologies

It is also possible to derive a dissimilarity from the graph of a subsumption relation across ontologies. The usual way to define subsumption *across ontologies* would be  $o \sqsubseteq o'$  iff  $o' \equiv o \cup o'$  ( $\cup$  here denoting the union of the axioms). We may also base it, in the continuity of the previous section on named class equivalence, i.e. an ontology subsumes another if it contains at least one equivalent class for each class of the other. Formally:

$$o \sqsubseteq o' \text{ if } \forall c' \in o', \exists c \in o; o \cup o' \models c \equiv c'$$

Such a relation can structure  $O$  in a graph  $\langle O, \sqsubseteq \rangle$  as the one featured at the centre of Figure 3.

From this graph, the dissimilarity between two ontologies may be the length of the shortest monotonous subsumption chain within the transitively reduced version of  $\sqsubseteq$  normalised by the longest such path +1 (with the measure set to 1 when no such path exists). The result is given, for ontologies of Figure 2, on Figure 3 (left).

## 4.4. Semantic categories of ontology

A way to define categories from a set of ontologies is to group equivalent ontologies in one single category. Different predicates ( $\equiv$ ) may be used to describes what counts as same ontologies. This may be based on strict syntactic equivalence, on named class equivalence, on equality of the sets of models, etc. This predicate allows us to group ontologies, but may be used for other purposes.

Finally, the dissimilarity defined between ontologies may be applied to the categories themselves. If they are compatible with the way to group ontologies into categories, i.e.  $\forall o, o', o'' \in O, o \equiv o' \Rightarrow d(o, o'') = d(o', o'')$ , then the dissimilarities defined above between two categories is the dissimilarity between any members of these categories.

This will be useful to integrate the dissimilarity  $d$  between ontologies in more sophisticated diversity measures (Section 6).

So far, we have defined simple semantic measures between ontologies and have provided a simple way to group ontologies. This does not tell us what the diversity of a set of agents is.

## 5. Measuring diversity as a dissimilarity

It is possible to consider that diversity can be measured by how far apart are the agents' knowledge. The simplest measure would be to consider whether ontologies are different or not. This can be generalised by introducing a dissimilarity between these ontologies introducing gradation in the differences.

### 5.1. Knowledge diversity as different ontologies

The most basic way to assess the diversity of a population is to count the different ontologies agents have. This requires to be able to identify when these ontologies are different.

This is achieved with an equivalence predicate ( $\equiv$ ) which can be a simple predicate to express that two ontologies are the same ( $o \equiv o'$ ) or different ( $\neg(o \equiv o')$ ). This equality may be a syntactic or semantic equality, as discussed in Section 4.4. In fact, it can be any equivalence relation.

Predicates based on a dissimilarity  $d$  between ontologies and a threshold  $\theta$  such that any pair of ontologies below the threshold are considered the same ( $o \equiv o'$  iff  $d(o, o') \leq \theta$ ), is not an equivalence relation in general. These may be used a posteriori to determine clusters of ontologies.

In order to aggregate this difference at the population level, it is possible to simply compute the number of different ontologies with respect to the number of agents:

$$\delta_{\equiv}(A) = \frac{|\{f(a); a \in A\} / \equiv|}{|A|}$$

In case of syntactic equivalence, the measure is:

$$\delta_{=} (A) = \frac{|O_A|}{|A|}$$

Such an indicator  $\delta_{=}$ , will provide the highest diversity measure: as soon as two ontologies are different, even slightly, their difference will be counted maximally. A finer view of this may be to compute a dissimilarity between ontologies instead of using a Boolean predicate.

## 5.2. Knowledge diversity as how ontologies are different

It is possible to define a similarity or dissimilarity between agents' ontologies measuring how alike or different they are. Here we consider a dissimilarity  $d : \mathcal{O} \times \mathcal{O} \rightarrow \mathbb{R}$  because it is compatible with the boolean measure. It will return 0 when the ontologies are the same.

Any dissimilarity (or similarity) between ontologies can be used, such as those presented in Section 4 and those mentioned in Section 3.

With respect to a population  $A$ , diversity may be assessed by agregating the dissimilarity between their ontologies. This may be achieved with:

- the average dissimilarity  $\delta_\alpha^d(A)$  in the population (works as well with predicates);
- the median  $\delta_\mu^d(A)$  of the dissimilarities (the set of dissimilarities is a multiset);
- the span or diameter of the population, i.e. the largest dissimilarity:  $\delta_\emptyset^d(A)$ .

Table 1 shows results for distributions of ontologies. It can be observed that the diameter and median are not very discriminative of different case. Similarly, average dissimilarity may return the same value for quite different cases: there are many ways to distribute ontologies at the same dissimilarity. Hence, it is useful to consider that for the same average dissimilarity, a lower the standard deviation means that this dissimilarity is regularly shared and the population is more diverse.

We extend this line of reasoning to taking distribution of knowledge in different categories instead of considering ontologies one by one.

## 6. Measuring diversity on a distribution

So far, we only considered dissimilarities between individual knowledge (or ontologies). When the set of possible objects becomes larger, it is customary to group them into categories. Indeed, each individual is usually different from the others but what counts as diversity is how specific categories of individuals are represented. These categories group objects (ontologies) whose difference is considered as insufficient to consider them different (hence diverse).

There are two ways to determine such categories:

**a priori** diversity is measured based on a set of predefined categories that can be assigned to individuals. These categories are independent from the data.

**a posteriori** the categories are determined through how individuals may be grouped together. This may be determined through clustering or applying factorial analysis to the ontologies. Such categories are usually relative to the considered data.

We will work on a priori categories because it allows to define what kind of diversity is to be *measured* (knowing what to look for) and not diversity that has to be *discovered* (returning what is the most diverse). For that purpose we will use a set  $K$  of categories. Each ontology belongs to one and only one category. The number of agents  $a \in A$  whose ontology  $f(a)$  belong to a category  $k \in K$  is denoted by  $\#k$ . Figure 1 shows the diversity of possible distributions of 10 objects in 5 categories presented along a linear order.

For the sake of simplicity, we will consider the categories of Section 4.4. They have the advantage that any ontology in these category may be taken as representing it, and, as already discussed, its dissimilarity to another category will be the same as its dissimilarity to any element of that other category. Hence, the set of category  $K$  will be a set of ontologies  $O$  and the distribution of agents' knowledge in this set will simply be  $O_A$ .

A measure of diversity in categories must take into account:

- the number of filled categories, which correspond to the variety of [14] (Figure 1),
- the distribution of objects in these categories, which correspond to separation (or intensity),
- how far appart are such categories, which corresponds to the disparity (or spread), and which can be measured by a dissimilarity.

We consider two families of measures in this context.

### 6.1. Weighted average dissimilarity in a distribution

A non-structured distribution is one in which the set of categories is simply a set with no additional structure. In such a case, two ontology within the same category will be counted as the same (0) and two objects in different category will be counted as different (1). The diversity measure may be computed with respect to the abundance of each category as:

$$\delta_{\equiv}(A) = \frac{\sum_{o \in O_A} \#o \times (|A| - \#o)}{|A| \times (|A| - 1)}$$

A structured distribution is defined when there exists a structure among the categories. This is given by providing a dissimilarity  $d$  between categories representing this structure.

The diversity is then the average dissimilarity between the objects' categories:

$$\delta_{\alpha}^d(A) = \frac{\sum_{o \in O_A} \#o \times (\sum_{o' \in O_A, o' \neq o} \#o' \times d(o, o'))}{|A| \times (|A| - 1)}$$

This is the same measure as in Section 5.2 but applied to categories instead of individual ontologies.

The structure may have various interpretations:

- The non-structured case may be obtained by taking the dissimilarity as the inverse of the identity matrix (0 on the diagonal, 1 otherwise).
- The linear structure that corresponds to the alignment of the five categories in Figure 1. The dissimilarity corresponds to how many slots the two categories are appart.
- Any dissimilarity between categories may be used and it can be based on syntactic or semantic cues. We will use the two semantic dissimilarities presented in Section 4.2 and 4.3.



## 6.2. Entropy of a distribution

The distribution among classes can be considered as the probability  $\frac{\#o}{|A|}$  of obtaining an ontology of category  $o$  when drawing it randomly among  $A$ , also called *relative abundance*. The entropy measures how much random this probability is, i.e. how much the category of a random individual may be predicted. This seems a very good candidate to measure diversity and it has been used for that purpose [11].

On such a distribution, a parametric measure of diversity can be defined [10, 21]:

$$\delta_q(A) = \left( \sum_{o \in O_A} \left( \frac{\#o}{|A|} \right)^q \right)^{\frac{1}{1-q}}$$

It is the exponent of a general entropy measure. This measure has been extended to structured distributions in which a similarity exists [22]. Such a similarity  $s$  may be obtained from a dissimilarity  $d$  by taking  $s(o, o') = e^{-d(o, o')}$ .

$$\delta_q^d(A) = \left( \sum_{o \in O_A} \frac{\#o}{|A|} \times \left( \sum_{o' \in O_A} \frac{\#o'}{|A|} \times e^{-d(o, o')} \right)^{q-1} \right)^{\frac{1}{1-q}}$$

The same dissimilarities between categories as above may be retained.

Depending of the parameter  $q$ , called the *order of diversity*, a different measure is obtained ranging from assigning more weight to the rarest category to assigning more weight to the commonest category, with more balanced weighting around 1. Here we use  $q = 2$  which corresponds to the inverse of the Gini-Simpson measure. The notion of ‘diversity profile’ (the evolution of diversity with  $q$ ) may be useful in defining a partial order between populations: a population  $A$  is more diverse than another  $A'$ ,  $A' \preceq A$  if for any  $q$ ,  $\delta_q^d(A) \geq \delta_q^d(A')$ .

This diversity measure is not normalised. It ranges within  $[1.0 + \infty]$ . When comparing a finite set of distributions, it is possible to normalise it as:

$$\bar{\delta}_q^d(A) = \frac{\delta_q^d(A) - 1}{\delta_q^d - 1}$$

with  $\delta_q^d$  the maximum value among the set.

Contrary to average dissimilarities, diameter and entropy are independent from scale (adding more agents in the same distribution will return the same measure).

Table 1 provides the measure values for these different distributions. Concerning the semantic measures, the five categories are considered as corresponding from left to right to the ontologies (A–E) of Figure 2.

As expected the diversity is minimum for distribution (a) for all measures. We also expect it to be maximum for distribution (g). But this did not happen in the case of the linear distance in which (e) has the highest diversity value. Actually (e) is polarised with two groups very far apart. However it is not very diverse as only two categories are filled.

distribution	(a)	(b)	(c)	(d)	(e)	(f)	(g)
Stats	$ A $	10	10	10	10	10	10
	$ O_A $	1	2	3	5	2	5
	$\frac{ O_A }{ A }$	.1	.2	.3	.5	.2	.5
Non struct.	$\emptyset$	0	1	1	1	1	1
	$\delta_{\mu}^d$	0.0	1.0	1.0	1.0	1.0	1.0
	$\delta_{\alpha}^d$	0	.56	.62	.67	.56	.82
	$\bar{\delta}_2^d$	0.0	.45	.54	.60	.45	.86
Linear	$\emptyset$	0	2	2	4	4	4
	$\delta_{\mu}^d$	0.0	2.0	1.0	1.0	4.0	1.0
	$\delta_{\alpha}^d$	0.0	1.11	.71	1.11	2.22	1.33
	$\bar{\delta}_2^d$	0.0	.43	.33	.48	.54	.70
Graph-B.	$\emptyset$	0	1	1	1	1	1
	$\delta_{\mu}^d$	0.0	1.0	.33	.33	1.0	.33
	$\delta_{\alpha}^d$	0.0	.56	.27	.37	.56	.47
	$\bar{\delta}_2^d$	0.0	.78	.39	.56	.78	.74
Name-b.	$\emptyset$	0	.5	.5	.5	.71	.29
	$\delta_{\mu}^d$	0.0	.5	.5	.5	.29	.5
	$\delta_{\alpha}^d$	0.0	.28	.31	.38	.16	.44
	$\bar{\delta}_2^d$	0.0	.52	.61	.74	.30	.94

**Table 1**

Measures corresponding to the distributions of Figure 1 with non structured, linearly structured (with the A-B-C-D-E order), graph-based and named-class-based semantic categories. In each case, is displayed the diameter, the median, the average dissimilarity and the normalised entropy-based diversity ( $q = 2$ ).

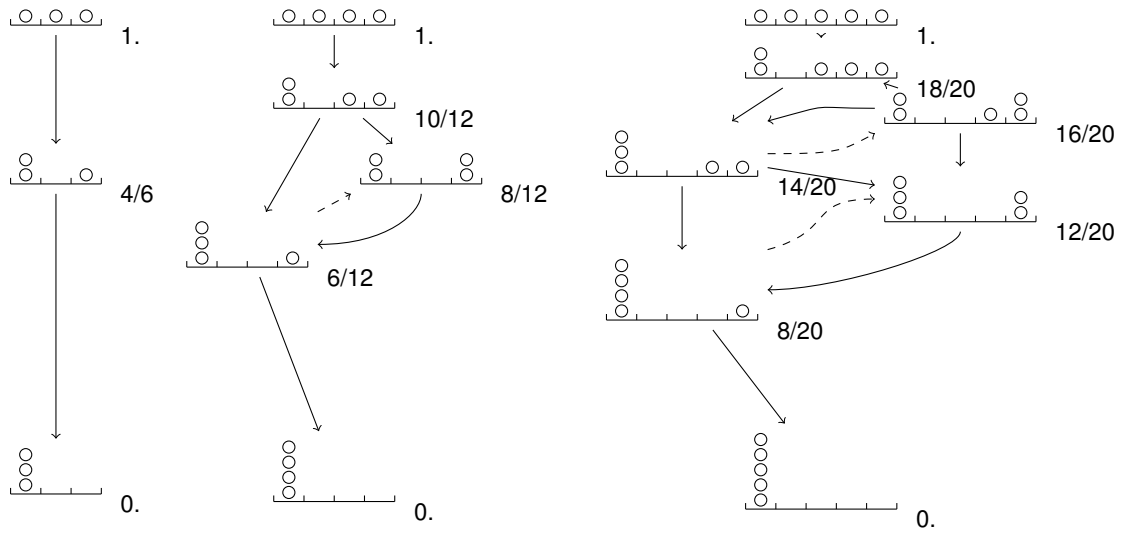
The entropy-based diversity measures seems to be behaving better. They are also very sensitive to the influence of the dissimilarity: there is no consensus between the orders induced by the measures except for the extrema. For instance, the graph-based semantic diversity has the same high values for (b) and (e) because the two categories which are filled have the same dissimilarity. These measures also do not account very well for the empty categories: they have been designed for applications for which most categories have some representations. Our example is likely too small.

In conclusion, it seems that entropy-based measures, by putting emphasis on equal dispersion of objects in all categories is a good measure of diversity, and that a semantic dissimilarity structures knowledge categories appropriately. However, it is not always the case that agents ontologies present themselves as distributions. In this case, a semantic dissimilarity alone should provide a good idea of diversity.

## 7. Controlling diversity

When one wants to observe experimentally the consequences of diversity, it is necessary to be able to control it.

The conceptually most straightforward way to achieve this is to determine the diversity measure



**Figure 4:** Patterns of ontology substitution for  $|A| = 3, 4, 5$ . Dashed transitions do not decrease diversity.

to be used and to design agent populations and their knowledge which comply to different values of these measures. This may be quite difficult because we already have two factors (agents and knowledge) and knowledge itself may be diverse in a variety of ways. Synthesising artificial ontologies is possible, but may result in non representative results.

In our experiments [5], we observed that some factors, e.g. number of object descriptors, have an influence on the diversity of the resulting knowledge. It would thus be tempting to control these factors in order to obtain the desired diversity. However, this is not a good way to control diversity because it would be very difficult to determine if the observed effects will be due to diversity or to these factors themselves.

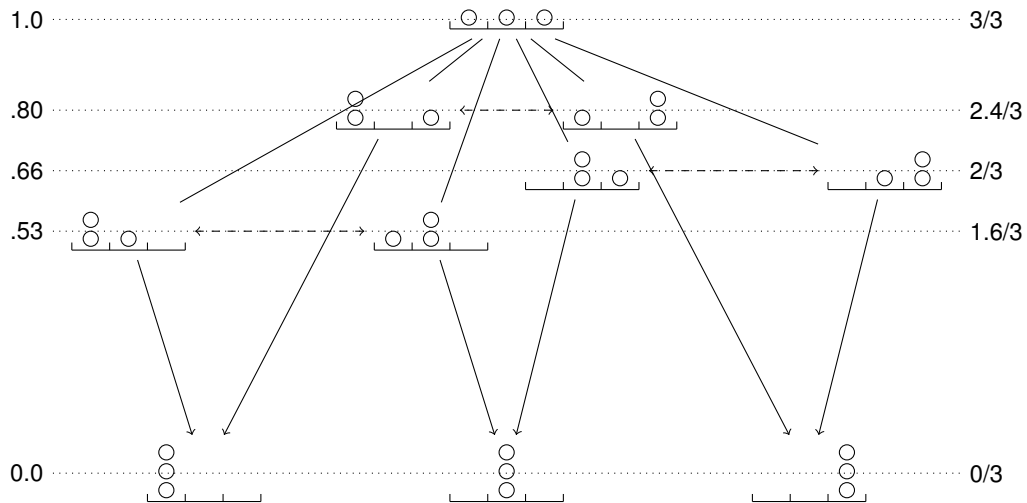
One possible way to improve on that is to perform such an experiment and to stop the simulation. What is obtained is a population of agents  $A$  with a ‘roof’ diversity  $\delta = \delta(A)$ . At that stage, if each agent knowledge is replaced by the ontology of one specific agent, taken at random or as the fittest knowledge for some measure, yielding distribution  $A'$ , then  $\delta(A')$  will be minimal. One further advantage of this procedure is that it compares distributions over the same set of agents and the same set of ontologies. Hence all distributional measures may be used and may be normalised without inconvenient. It is thus possible to obtain different distributions with diversity ranging in  $[0.. \delta]$  by distributing not 1 but  $n \in [1..|A|]$  different ontologies among the agents.

## 7.1. Boolean approach

In the case of a boolean distance function, the natural approach is to replace an ontology by another (already used by one agent).

The diversity of a (multi)set of such ontologies is computed by  $\delta(A)$  (with any of the  $\delta$  defined before).

We consider a simple operation  $r$  which consists of replacing one ontology by another ontology



**Figure 5:** The different distributions of three ontologies in three categories and the possible transitions for  $r$ . Dashed transitions do not decrease diversity.

that is already present in the multiset. Applying iteratively this operation progressively reduces the diversity in the multiset. These successive applications define a path in the graphs of Figure 4. If one takes care to never apply an operation that increases diversity, then this path goes from the top to (one of) the bottom(s).

Figure 4 presents multisets of ontologies of the same cardinality, up to category permutation, related by arrows representing the application of  $r$ ; dashed arrows are those which do not decrease diversity. This enable to define different levels of diversity that can be used for the experiments. For the sake of simplicity, diversity is here measured by average dissimilarity in a non structured set of categories. The more ontologies are initially in the set, the more levels can be defined.

## 7.2. Simple dissimilarity-based approach

It may be more interesting to control diversity based on non-boolean dissimilarities between ontologies. The same type of reasoning may be applied with  $\delta^d$ .

Hence, contrary to what appears in Figure 4, the levels do not depend on the number of occurrences of ontologies only. Their respective dissimilarity measures will play an important role and lead to a larger variety of diversity levels.

Consider three ontologies  $A$ ,  $B$ , and  $C$  such that  $d(A, B) = .8$ ,  $d(B, C) = 1$ . and  $d(A, C) = 1.2^1$ . These may be considered as ontologies  $o \in O$  or categories  $k \in K$ . The development of the pattern above on these three ontologies provide the result of Figure 5.

The opportunities are here very limited due to the strict pattern followed for only three ontologies by only three agents. Starting with four, the diagram becomes largely more complex with more paths from top to bottom.

Figure 4 and 5 show that the diversity levels are fixed and relatively more dense over .5 than

<sup>1</sup>This averages to 1., but there is no obligation. Moreover, it still provides high level of diversity (all over 0.5).

below it. This may be a problem when one wants to control better the given level. However, this is also very dependent on the diversity measure, hence unless one knows very precisely the meaning of these levels, it may be better to rely only on the order between distributions. Moreover, on a statistical basis if the goal is to establish the relation between a similarity measure and another measure, it is not strictly necessary to have a regularly spaced sample.

A simple algorithm for generating new sets of ontologies consists of computing the average dissimilarity between each ontologies and all the others. Then replacing the one with the higher dissimilarity with either (a) the most central (closer to the barycentre), or (b) the most central with respect to the remaining ones. Such an algorithm would again define a path within the lattice of possible ontology distributions.

Increasing the ratio agent/ontologies and using more sophisticated measures, such as the entropy-based ones, provide a reasonable way to collect samples of various diversity-levels.

### 7.3. Finely controlling diversity

The algorithm sketched in the previous section still does not allow to control finely the obtained levels. If someone would like to obtain a set of ontologies for 1., .75, .5, .25 and 0. diversity. It will not return exactly this. Worse, it will not return the best possible solution. It is possible to express it as an optimisation problem. Indeed, the goal is to find, for each diversity level  $l$ , the assignment  $\#$  that provides the multiset  $O^*$  whose diversity  $\delta(O)$  is the closest to  $l$ , i.e.

$$O^* = ARGMIN_{\#} |l - \delta(O)|$$

under the constraints that:

$$\forall o \in O, \#o \geq 0 \text{ and } \sum_{o \in O} \#o = |A|$$

A very expensive algorithm for achieving this would be to compute all the replacement combinations, to measure their diversity, and to retain those which are the closest to the expected levels.

## 8. Conclusions

This work was motivated by the need to measure and control knowledge diversity. It seems to us that the approach consisting in taking advantage of measures already defined to assess diversity and adjoining them with measures already defined to assess knowledge difference is a promising one. However, there are so many such measures to be combined that a careful examination of the properties of these and their combinations is an interesting perspective. This would benefit from carefully studying the interactions between axioms governing both components in the spirit of [15, 11]. We briefly discussed one practical way to control knowledge diversity for the purpose of experiments. This is a critical issue that has no simple solution. However, the problem has to be tied to the knowledge diversity measure.

## Code and data availability

All measures and algorithms have been implemented in Python and can be retrieved from <https://moex.inria.fr/software/kdiv/>.

## Acknowledgments

This work has been partially supported by the MIAI Knowledge communication and evolution chair (ANR-19-P3IA-0003). We thank the reviewers for their careful reading and for helping clarify some formulations.

## References

- [1] A. Stirling, A general framework for analysing diversity in science, technology and society, *Journal of the royal society—Interface* 4 (2007) 707–719.
- [2] L. Hong, S. Page, Groups of diverse problem solvers can outperform groups of high-ability problem solvers, *Proceedings of the national academy of sciences* 101 (46) (2004) 16385–16389.
- [3] D. Noble, M. Prates, D. Bossle, L. Lamb, Collaboration in Social Problem-Solving: When Diversity Trumps Network Efficiency, in: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, AAAI’15*, AAAI Press, 1277–1283, 2015.
- [4] D. Reed, R. Frankham, Population fitness is correlated with genetic diversity, *Conservation biology* 17 (2003) 230–237.
- [5] Y. Bourahla, M. Atencia, J. Euzenat, Knowledge improvement and diversity under interaction-driven adaptation of learned ontologies, in: U. Endriss, A. Nowé, F. Dignum, A. Lomuscio (Eds.), *Proc. 20<sup>th</sup> ACM international conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, London (UK), 242–250, URL <http://www.ifaamas.org/Proceedings/aamas2021/pdfs/p242.pdf>, 2021.
- [6] J. Goguen, Support for ontological diversity and evolution, URL <https://cseweb.ucsd.edu/~goguen/papers/onto-intgn.html>, 2005.
- [7] F. Baader, D. Calvanese, D. McGuinness, D. Nardi, P. Patel-Schneider, *The description logic handbook: Theory, implementation and applications*, Cambridge University Press, 2 edn., 2007.
- [8] C. Aremu, Exploring statistical tools in measuring genetic diversity for crop improvement, in: M. Çalışkan (Ed.), *Genetic diversity in plants*, chap. 17, IntechOpen, 339–348, 2012.
- [9] M. Avolio, J. Beaulieu, E. Lo, M. Smith, Measuring genetic diversity in ecological studies, *Plant ecology* 213 (17) (2012) 1105–1115.
- [10] L. Jost, Entropy and diversity, *Oikos* 113 (2) (2006) 363–375.
- [11] T. Leinster, *Entropy and diversity: the axiomatic approach*, Cambridge university press, ISBN 9781108965576, URL <https://arxiv.org/pdf/2012.02113.pdf>, 2021.
- [12] E. Cheng, How to measure diversity —mathematical theory gives some rigor to discussion of a sensitive social and political issue, *Wall street journal* 2017.

- [13] H. Cooke, I. Keppo, S. Wolf, Diversity in theory and practice: a review with application to the evolution of renewable energy generation in the UK, *Energy Policy* 61 (2013) 88–95.
- [14] D. Harrison, K. Klein, What’s the difference? Diversity constructs as separation, variety, or disparity in organizations, *Academy of management review* 32 (2007) 1199–1228.
- [15] P. Ramaciotti Morales, R. Lamarche-Perrin, R. Fournier-S’niehotta, R. Poulain, L. Tabourier, F. Tarissan, Measuring diversity in heterogeneous information networks, *Theoretical computer science* 859 (2021) 80–115, URL [https://pedroramaciotti.github.io/files/publications/2021\\_TCS.pdf](https://pedroramaciotti.github.io/files/publications/2021_TCS.pdf).
- [16] F. Giunchiglia, M. Fumagalli, On knowledge diversity, in: Proc. 4th International Workshop on Ontology Modularity, Contextuality, and Evolution (WOMoCoE), Graz (AT), URL <http://ceur-ws.org/Vol-2518/paper-WOMOCOE2.pdf>, 2010.
- [17] J. David, J. Euzenat, Comparison between ontology distances (preliminary results), in: Proc. 7th conference on international semantic web conference (ISWC), Karlsruhe (DE), vol. 5318 of *Lecture notes in computer science*, 245–260, URL <http://www.springerlink.com/content/cj22428300485784/>, 2008.
- [18] A. Mädche, S. Staab, Measuring Similarity between Ontologies, in: Proc. 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW), vol. 2473 of *Lecture notes in computer science*, Siguenza (ES), 251–263, 2002.
- [19] J. David, J. Euzenat, O. Sváb-Zamazal, Ontology similarity in the alignment space, in: Proc. 9th international semantic web conference (ISWC), Shanghai (CN), 129–144, URL <https://exmo.inria.fr/files/publications/david2010b.pdf>, 2010.
- [20] J. Euzenat, C. Allocca, J. David, M. d’Aquin, C. Le Duc, O. Sváb-Zamazal, Ontology distances for contextualisation, Tech. Rep. 3.3.4, NeOn, URL <https://exmo.inria.fr/files/reports/neon-334.pdf>, 2009.
- [21] M. Hill, Diversity and evenness: a unifying notation and its consequences, *Ecology* 54 (2) (1973) 427–432, URL [https://pedroramaciotti.github.io/files/publications/2021\\_TCS.pdf](https://pedroramaciotti.github.io/files/publications/2021_TCS.pdf).
- [22] T. Leinster, C. Cobbold, Measuring diversity: the importance of species similarity, *Ecology* 93 (3) (2012) 477–489, URL <https://www.maths.gla.ac.uk/~cc/pdf/Leinster2011.pdf>.