Rough Set Semantics for Identity on the Web

Wouter Beek and **Stefan Schlobach** and **Frank van Harmelen**
Vrije Universiteit Amsterdam
De Boelelaan 1081a
1081HV Amsterdam
The Netherlands

Abstract

Identity relations are at the foundation of many logic-based knowledge representations. We argue that the traditional notion of equality, is unsuited for many realistic knowledge representation settings. The classical interpretation of equality is too strong when the equality statements are re-used outside their original context. On the Semantic Web, equality statements are used to interlink multiple descriptions of the same object, using `owl:sameAs` assertions. And indeed, many practical uses of `owl:sameAs` are known to violate the formal Leibniz-style semantics.

We provide a more flexible semantics to identity by assigning meaning to the subrelations of an identity relation in terms of the predicates that are used in a knowledge-base. Using those indiscernability-predicates, we define upper and lower approximations of equality in the style of rought-set theory, resulting in a quality-measure for identity relations.

1 Introduction

Identity relations are a cornerstone of logic-based knowledge representation. They allow to state and relate properties of an object using multiple names for that object, and conversely, they allow to infer that different names actually refer to the same object.

Identity relations are at the foundation of the Linked Open Data initiative and the Semantic Web (SW) in general. The SW consists of sets of assertions that are published on the Web by different authors operating in different contexts, often using different names for the same object. Identity relations allow the interlinking of these multiple descriptions of the same thing.

Identity is often understood as the sharing of all properties between two objects with different names (principle of indiscernibility). In the SW this traditional notion of identity is expressed by the `owl:sameAs` property.
According to the traditional semantics of the identity relation, identical terms can be replaced for one another in all non-modal contexts *salva veritate*. Practical uses of `owl:sameAs` are known to violate this strict condition (Halpin et al. 2010a; 2010b).

On the SW, identity assertions are extra strong because of the Open World Assumption. Stating that two objects are the same implies that from now on no new property can be stated about only one of those objects Moreover, whether or not two objects share the absence of a property cannot be concluded based on the absence of a property assertion.

Improving on the existing semantics of identity, we have the following research goals:

1. In an identity relation the pairs all look the same. We want to characterize subrelations of an identity relation in terms of the predicates that are important in a particular context.
2. Based on an existing identity relation we want to give semantically motivated suggestions for extending or limiting the identity relation.
3. We want to assess the quality of an identity relation based on the consistency with which it is applied to the data.

2 Related work

Existing research suggests six different solutions for the problem of identity on the SW.

[1] Introduce weaker versions of `owl:sameAs` (Halpin et al. 2010a; McCusker and McGuinness 2010). Candidates for replacement are the SKOS concepts `skos:related` and `skos:exactMatch`. The former is not transitive, thereby limiting the possibilities for reasoning. The latter is transitive, but can only be used in certain contexts. It is not defined in what contexts it can be used.

[2] Restrict the applicability of identity relations to specific contexts. In terms of SW technology, identities are expected to hold within a named graph or within a namespace, but not necessarily outside of it (Halpin et al. 2010a). (de Melo 2013) has successfully used the Unique Names Assumption within namespaces in order to identify many (arguably) spurious identity statements.

[3] Introduce additional vocabulary that does not weaken but extend the existing identity relation. (Halpin et al. 2010a) mention an explicit distinction that can be made between mentioning a term and using a term. Other possible extensions of `owl:sameAs` take the Fuzzyness and/or uncertainty of identity statements into account.

[4] Use domain-specific identity relations (McCusker and McGuinness 2010). Such domain-specific links are only locally valid, thereby limiting knowledge reuse.

[5] Change the modeling practice (Halpin et al. 2010a; Ding et al. 2010a). Introducing checks on editing operations

violates one of the fundamental underpinnings of the SW: that anybody is allowed to say anything about anything (Antoniou et al. 2012).

(Ding et al. 2010b) who show that network analysis of the occurrence of `owl:sameAs` in datasets can provide insights into the ways in which identity is used. These latter endeavors have not yet been related to the semantics of identity.

3 Approach

First we give a short outline of our approach and then we provide a more detailed description of the individual steps.

Outline of the approach

We start by assuming that we are given an identity relation $\approx$. We will then reinterpret this relation as an indiscernibility relation relative to different sets of predicates. Pairs that have the same indiscernibility predicates are simi-discernible, i.e.: they discern resources based on the same criteria. Simi-discernibility is an equivalence relation which partitions all pairs and thus also the identity relation $\approx$. The members of the indiscernibility partition have a certain overlap with the original identity relation. The overlap between an indiscernibility subset and the identity relation is called an *identity subrelation*. Each identity subrelation is characterized in terms of predicates from the domain vocabulary. Different forms of identity can therefore be distinguished and meaningfully described. Based on whether there is a complete or a partial overlap between the simi-discernible partition members and the identity subrelations, these partition members belong either to the lower ($\underline{\approx}$) or to the higher approximation ($\overline{\approx}$) of $\approx$. Besides setting a lower and a higher bound to the identity relation, we can also calculate the quality of the identity relation and the precision of each identity subrelation.

Preliminaries

G denotes an RDF graph. It consists of a set of ground binary predicates $p(s, o)$, called "triples" in SW jargon, and often written as $\langle s, p, o\rangle$. These triples form a graph with all subjects s and objects o as nodes, and each assertion $p(s, o)$ corresponds to a directed edge labelled p between s and o.

We identify subsets of RDF terms based on their positional occurrence in triples in G: S_G, P_G, and O_G denote the subject, predicate and object terms in G respectively.

The interpretation I maps RDF terms onto resources, and triples onto truth values. The extension function Ext maps resources onto pairs of resources. $I(\langle s, p, o\rangle)$ is true iff $\langle I(s), I(o)\rangle \in Ext(I(p))$.

Reinterpreting identity as indiscernibility

We start by assuming that we are given an identity relation $\approx$, which partitions the subject terms S_G according to equation 1.

$$[x]_{\approx} = \{y \in S_G \mid x \approx y\} \tag{1}$$

Identity can be defined as the smallest equivalence relation, i.e. the most fine-grained partition of S_G. For reasoning purposes, the fact that $\approx$ is an equivalence relation is important, allowing symmetrical and transitive inferences. Identity implies indiscernibility with respect to all properties.

We can generalize the notion of indiscernibility by parameterizing the set of properties with respect to which indiscernibility is determined. According to this generalization, resources x and y are indiscernible with respect to a set of properties $PO \subseteq P_G \times O_G$ iff $\forall po \in PO(po(x) \leftrightarrow po(y))$ is the case. Every indiscernibility relation is also an equivalence relation, although not necessarily the smallest one. Moreover, every indiscernibility relation defined over the domain S_G is also an identity relation, but over a different domain.

We now reinterpret the identity relation $\approx$ as if it were an indiscernibility relation whose set of properties PO is implicit. Based on the extensional specification of the identity relation, we make the set of properties with respect to which it is indiscernible explicit. Definition 1 makes explicit the properties relative to which the terms x_i are indiscernibile.

Definition 1 (Indiscernibility properties).

$$\mathbb{IND}\text{-}\mathbb{PO}_{\approx}(\{x_1, \ldots, x_n\}) = \{\langle p, o\rangle \in P_G \times O_G \mid \bigwedge_{1 \le i \le n} \exists p_i \in [p]_{\approx}, \exists o_i \in [o]_{\approx}(\langle x_i, p_i, o_i\rangle \in G)\}$$

Notice that in definition 1 we close both the predicate terms p and the object terms o under identity. Performing these closures is important in order to identify the relevant indiscernibility properties.

In the above, we were interested in the properties that resources share with one other. But we are also interested in the predicates that are shared by a set of resources. This amounts to a simple abstraction of definition 1, equating the sets of objects (closed under identity) and only returning the set of shared RDF predicate terms (see definition 2).

Definition 2 (Indiscernibility predicates).

$$\mathbb{IND}\text{-}\mathbb{P}_{\approx}(\{x_1, \ldots, x_n\}) = \{p \in P_G \mid \exists_{p_1,\ldots,p_n \in [p]_{\approx}}([\{o \in O_G \mid \langle x_1, p_1, o\rangle\}]_{\approx} = \ldots = [\{o \in O_G \mid \langle x_n, p_n, o\rangle\}]_{\approx})\}$$

Discerning the same

In the previous section we saw that resources are *indiscernible* with respect to PO iff they cannot be told apart in a language that only contains the properties denoted by PO (the so-called indiscernibility properties):

In the same vein, and builing upon definition 2, we say that two pairs of resources are *simi-discernible* iff their indiscernibility predicates $P \subseteq P_G$ are the same.

When we look at the pairs that constitute (the extension of) an identity relation, all identity assertions look the same. But when we take the considerations of the previous section into account, we see that within a given identity relation there are pairs that assert indiscernibility based on different domain predicates. Stating this formally, simi-discernibility is an equivalence relation on pairs of resources, which induces a partition of the Cartesian product of the domain. Definition 3 makes this concrete in terms of the earlier definitions.

Definition 3 (Simi-discernibility relation).

$$\equiv_{\mathbb{IND}\text{-}\mathbb{P}_{\approx}} = \{\langle\langle x_1, x_2\rangle, \langle y_1, y_2\rangle\rangle \in (S_G^2)^2 \mid \mathbb{IND}\text{-}\mathbb{P}_{\approx}(\{x_1, x_2\}) = \mathbb{IND}\text{-}\mathbb{P}_{\approx}(\{y_1, y_2\})\}$$

Partitioning identity

The members of the partition induced by $\equiv_{\mathbb{IND}\text{-}\mathbb{P}_{\approx}}$ are sets of resource pairs that share the same sharing properties.

Notice that the partitioned pairs contain but are not limited to the identity pairs. Therefore, for sets of pairs closed under simi-discernibility we have the following three possibilities:

1. All pairs are identity pairs. This characterizes a consistent portion of the identity relation, since no simi-discernible pair is left out of this set.
2. Some pairs are identity pairs. This characterizes a portion of the identity relation which is not applied consistently with respect to the simi-discernibility relation.
3. No pairs are identity pairs. This characterizes a portion of the collection of pairs that is consistently kept out of the identity relation.

Each member of the simi-discernibility partition that is not of the third kind, i.e. every set of pairs that contains some identity pair, can be though of as an identity subrelation. The simi-discernibility partition also partitions the identity relation into *identity subrelations*. Each identity subrelation can be described in terms of its discernibility predicates, i.e. in meaningful terms drawn from the domain vocabulary.

Quality & Approximation

Not all identity subrelations have the same quality. Indeed, when we look at the subdivision into three 'categories' above, we are able to distinguish between a lower approximation of identity, as the union of subrelations from the first category (definition 4), and a higher approximation of identity, as the union of subrelations from both the first and the second category (definition 5).

Definition 4 (Lower approximation).

$$x \in \underline{\approx} \iff \{y \mid x \equiv_{\mathbb{IND}\text{-}\mathbb{P}_{\approx}} y\} \subseteq \approx$$

Definition 5 (Higher approximation).

$$x \in \overline{\approx} \iff \{y \mid x \equiv_{\mathbb{IND}\text{-}\mathbb{P}_{\approx}} y\} \cap \approx \neq \emptyset$$

Based on these approximations we can give the rough set representation $\langle \underline{\approx}, \overline{\approx} \rangle$ of identity relation $\approx$. The quality of a rough set representation is given in definition 6. The intuition behind this quality measure is that the crispness of a set should be proportional to the quality of the identity relation on which it is based. Since a consistently applied identity relation has relatively many partition sets that contain either no identity pairs (small value for $\underline{\approx}$) or only identity pairs (large value for $\overline{\approx}$), a more consistent identity relation has a higher accuracy.

Definition 6 (Quality).

$$\alpha(\approx) = \frac{|\underline{\approx}|}{|\overline{\approx}|}$$

4 Conclusion

In this paper we presented a new approach for characterizing, extending, retracting, and assessing identity relations. Our approach does this in purely qualitative terms, using schema semantics.

In section 1 we enumerated three research goals. The first goal is met, since an indiscernibility partition characterizes identity subrelations based on the predicates P (closed under identity) for which the pairs in that sets are indiscernible. In this way we can distinguish between different types of identity by treating P as a description of a (sub)set of identity pairs. We suggest that the meaning of an identity relation and its subrelations is partially defined in its use, i.e., in the indiscernibility criteria it embodies.

The second goal is met, since the notion of a rough set allows us to distinguish between pairs that must be (lower approximation) and those that may be (higher approximation) in the identity relation. If we want to add/remove pairs of the identity relation, we should not consider pairs of the former but only pairs of the latter kind.

The third goal is met, since the measure for rough set accuracy is based on the discernibility criteria of an identity set. The crispness of the set is proportional to the quality of the identity relation, based on its semantic consistency.

References

Antoniou, G.; Groth, P.; van Harmelen, F.; and Hoekstra, R. 2012. *A Semantic Web Primer (Third Edition)*. The MIT Press.

de Melo, G. 2013. Not quite the same: Identity constraints for the web of linked data. In *Proceedings of the American Association for Artificial Intelligence 2013*.

Ding, L.; Shinavier, J.; Finin, T.; and McGuinness, D. L. 2010a. owl:sameas and linked data: an empirical study.

Ding, L.; Shinavier, J.; Shangguan, Z.; and McGuinness, D. 2010b. SameAs networks and beyond: Analyzing deployment status and implications of `owl:sameAs` in linked data. In Patel-Schneider, P. F.; Pan, Y.; Hitzler, P.; Mika, P.; Zhang, L.; Pan, J.; Horrocks, I.; and Glimm, B., eds., *The Semantic Web – ISWC 2010*, volume 6496 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg. 145–160.

Halpin, H.; Hayes, P.; McCusker, J.; McGuinness, D.; and Thompson, H. 2010a. When `owl:sameAs` isn't the same: An analysis of identity in linked data. In Patel-Schneider, P. F.; Pan, Y.; Hitzler, P.; Mika, P.; Zhang, L.; Pan, J. Z.; Horrocks, I.; and Glimm, B., eds., *The Semantic Web – ISWC 2010*, volume 6496 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg. 305–320.

Halpin, H.; Hayes, P.; McCusker, J.; McGuinness, D.; and Thompson, H. S. 2010b. *When `owl:sameAs` Isn't the Same: An Analysis of Identity in Linked Data*. Lecture Notes in Computer Science. Springer Berlin / Heidelberg. 305–320.

McCusker, J., and McGuinness, D. 2010. Towards identity in linked data. *Proceedings of OWL Experiences and Directions Seventh Annual Workshop*.