A Contextualised Semantics for `owl:sameAs`

Wouter Beek[1], Stefan Schlobach[1], and Frank van Harmelen[1]

Dept. of Computer Science, VU University Amsterdam, NL
{w.g.j.beek,stefan.schlobach,frank}@vu.nl

Abstract. Identity relations are at the foundation of the Semantic Web and the Linked Data Cloud. In many instances the classical interpretation of identity is too strong for practical purposes. This is particularly the case when two entities are considered the same in some but not all contexts. Unfortunately, modeling the specific contexts in which an identity relation holds is cumbersome and, due to arbitrary reuse and the Open World Assumption, it is impossible to anticipate all contexts in which an entity will be used. We propose an alternative semantics for `owl:sameAs` that partitions the original relation into a hierarchy of subrelations. The subrelation to which an identity statement belongs depends on the dataset in which the statement occurs. Adding future assertions may change the subrelation to which an identity statement belongs, resulting in a context-dependent and non-monotonic semantics. We show that this more fine-grained semantics is better able to characterize the actual use of `owl:sameAs` as observed in Linked Open Datasets.

1 Introduction

Identity relations are at the foundation of the Semantic Web and the Linked Data initiative. 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. The Semantic Web 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. However, the traditional notion of identity expressed by `owl:sameAs` [17] is problematic when objects are considered the same in some contexts but not in others. According to the standard semantics, 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 [10,11]. The standing practice in such cases is to use weaker relations of relatedness such as `skos:related` [16]. Unfortunately, these relations suffer from the opposite problem of having almost no formal semantics, thereby limiting reasoners in drawing inferences. In this paper we introduce an alternative semantics for `owl:sameAs` that is parameterized over the particular properties that are taken into account when deciding on identity. This allows formally specified context-specific adaptations of the identity relation. We give the formal definition, provide working examples and present a small-scale implementation.

The rest of the paper is structured as follows. In the next section we analyze problems caused by the traditional notion of identity. After surveying existing work in Section 2 we present our approach in Section 5 and enumerate some of the applications of this new semantics in Section 6. We illustrate the results of applying our formalism to Linked Datasets in Section 7 based on a working implementation. Section 8 concluses.

2 Related work

Existing research suggests the following six solutions for the problem of identity.

Introduce weaker versions of `owl:sameAs` [10,15] Candidates for replacement are the SKOS concepts `skos:related` and `skos:exactMatch` [16]. The former is not transitive, thereby limiting the possibilities for reasoning. The latter is transitive but is said to only be used in certain contexts without stating what those contexts of use are [16]. As example we will quote the intended use of property `skos:exactMatch` according to the SKOS specification: "[exactMatch] is used to link two concepts, indicating a high degree of confidence that the concepts can be used interchangeably across a wide range of information retrieval applications." From this it follows that the meaning of some SKOS relations changes over time, as IR applications become more advanced. Another problem with using weaker notions such as relatedness, is that everything is related to everything in *some* way.

Restrict the applicability of identity relations to specific contexts In terms of Semantic Web technology, identities are expected to hold within a named graph or within a namespace but not necessarily outside of it [11]. Gerard de Melo [4] has successfully used the Unique Name Assumption within namespaces in order to identify many (arguably) spurious identity statements.

Introduce additional vocabulary that does not weaken but extend the existing identity relation. Halpin et al. [10] mentions an explicit distinction that could be made between mentioning a term and using a term, thereby distinguishing an object and a Web document describing that object. Other possible extensions of `owl:sameAs` may take the fuzziness and/or uncertainty of identity statements into account [14].

Use domain-specific identity relations [15] For instance "x and y have the same medical use" for identity in the domain of medicine and "x and y are the same molecule" for identity in the domain of chemistry. The downside to this solution is that domain-specific links are only locally valid, thereby limiting knowledge reuse.

Change modeling practice Possibly in a (semi-)automated way, by adapting visualization and modeling toolkits to produce notifications upon reading SW data or by posing additional restrictions on the creation and alteration of data. For example, adding an RDF link could require reciprocal confirmation from the maintainers of the authorities of the respective relata [11,5]. The problem with introducing checks on editing operations is that it violates one of the fundamental underpinnings of the SW according to which anybody is allowed to say anything about anything (AAA) [2].

Extract network properties of `owl:sameAs` datasets Ding et al. [6] shows that network analysis can provide insights into the ways in which identity is used on the Semantic Web. However, results from network analytics research have not yet been related to the semantics of the identity relation. We believe that utilizing network theoretic aspects in order to determine the meaning of identity statements may be interesting for future research.

What the existing approaches have in common is that many adaptations have to be made – introducing terminology, instructing modelers, converting datasets – in order to resolve only some of the problems of identity. Our approach provides a way of dealing with the heterogeneous real-world usage of identity in the Semantic Web that can be automated and that does not require changes to modeling practices or existing datasets.

Our work bears some resemblance to existing work on key discovery: the practice of finding sets of properties that allow subject terms to be distinguished [20]. In particular, our notion of indiscernibility properties (Definition 2) is identical to the notion of a key in key discovery.

3 Motivation

Entities that are the same share the same properties. This 'indiscernibility of identicals' (Principle 1) is attributed to Leibniz [7] and its converse, the 'identity of indiscernibles' (Principle 2), states that entities that share the same properties are the same. Ψ denotes the set of all properties.

Principle 1 (Indiscernibility of identicals)**.** $a = b \rightarrow (\forall\,\phi \in \Psi)(\phi(a) = \phi(b))$

Principle 2 (Identity of indiscernibles)**.** $(\forall\,\phi \in \Psi)(\phi(a) = \phi(b)) \rightarrow a = b$

Although Principles 1 and 2 provide necessary and sufficient conditions for identity, they do not point towards an effective procedure for enumerating the extension of the identity relation. Moreover, the principle is circular since $a = b$ implies that a and b share the properties "$= a$" and "$= b$". Even though this principle does not allow a positive identification of identity pairs, it does provide an exclusion criterion; namely objects that are known to not share some property are also known to not be identical.

Identity poses several problems that are not specific to the SW. Firstly, identity does not hold across (all) modal contexts, allowing Lois Lane to believe that Superman saved her without requiring her to believe that Clark Kent saved her. Secondly, identity is context-dependent [9]. For instance, two medicines may be considered the same in terms of their chemical substance while not being considered the same commercial drug (e.g., because they are produced by different companies). Thirdly, identity over time poses problems since a ship may still be considered the same ship, even though all its original components have been replaced by new ones [13]. Lastly, there is the problem of identity under counterfactual assertions, that allow *any* property of an individual to be negated [12]. E.g., "If my parents would not have met then I would not have been born." These four problems indicate that a real-world semantics of identity should be context-dependent and non-monotonic.

Besides the generic problems of identity there are problems that are specific to the Semantic Web and its particular semantics and pragmatics. The OWL semantics for identity is given in Definition 1, where $\mathcal{I}$ is the interpretation function mapping terms to resources and EXT is the extension function mapping properties to pairs of resources.

Definition 1 (Semantics of `owl:sameAs`).

$$\langle \mathcal{I}(a), \mathcal{I}(b) \rangle \in \mathit{EXT}(\mathcal{I}(\texttt{owl:sameAs})) \Leftrightarrow \mathcal{I}(a) = \mathcal{I}(b)$$

Notice that Definition 1 defines `owl:sameAs` in terms of the identity relation '$=$' that we have previously argued to be highly problematic. Identity assertions are extra strong on the Semantic Web because of the Open World Assumption. Stating that two entities are the same implies that from now on no new property can be stated about only one of those entities. This follows from Definition 1 in combination with the principle of substitutivity *salva veritate*. For instance, if one source asserts that medicines b and c are the same based on them having the same chemical composition, this prohibits a future source from stating that b and c are produced by different companies, without resulting in an inconsistent state. In other words: every identity assertion makes a very strong claim that quantifies over the entire set Ψ (see Principles 1 and 2). Moreover, on the Semantic Web the set of properties Ψ is constantly increasing. In fact, since an RDF property has both in- and extension, the number of properties is not even limited by the size of the universe of discource, as different properties may have the same extension. Finally, whether or not two objects share the absence of a property, i.e., a property of the form "does not have the property ϕ", cannot be concluded based on the absence of a property assertion. Such 'negative knowledge' must be provided explicitly using, e.g., class restrictions. All this amounts to saying that *there can in principle not be an effective procedure for establishing the truth of `owl:sameAs` assertions*. (Establishing the falsehood of such assertions is of course possible, see our comments above.)

When we take the social component of the Semantic Web into account as well, we observe that modelers sometimes have different opinions about whether two objects are identical or not. While in some cases this may be due to a difference in modeling competence, there is also the more fundamental problem that two modelers may be constructing (parts of) the same knowledge base from within different contexts. Since Semantic Web knowledge is intended to be re-used in unanticipated contexts, the presence of knowledge from different perspectives is one of its inherent characteristics. In addition, the term `owl:sameAs` is overloaded to not only denote the *semantics* of identity but also the *practice* of linking datasets together. The fifth star of Linked Open Data (LOD) publishing [3] states that you should "Link your data to other people's data to provide context," and data is almost exclusively linked using the `owl:sameAs` property [1]. Concluding, from the social point of view today's requirements on Semantic Web modelers are unreasonably high when they are required to anticipate future additions by others while asserting identity links in accordance with the strict semantics. At the same time Linked Data best practices state that modelers should make those links in order to contextualize their knowledge.

Based on the above analysis, we can state the following desiderata for a semantics of identity that does not suffer from the identified problems:

1. The uniform identity relation should be reinterpreted in multiple subrelations that should be characterized in terms of the contexts in which those subrelations appear.
2. An alternative semantics for identity should be able to derive entailment results with respect to a given context.
3. Based on an existing identity relation, semantically motivated feedback should be given to the modeler about the different context-dependent subrelations that are currently expressed.
4. The quality of an identity relation should be quantified in terms of the consistency with which its context-dependent subrelations are applied to the data. Specifically, suggestions for extending or limiting the identity subrelations should be derived by automated means.

4 Preliminaries

Here we introduce the terminology and symbolism that is used throughout the rest of this paper.

RDF syntax RDF terms (RDF_T) come in three flavors: blank nodes (RDF_B), IRIs (RDF_I) and literals (RDF_L). Statements in RDF are triples $\langle s, p, o \rangle$ that are members of $(RDF_B \cup RDF_I) \times RDF_I \times RDF_T$. In a triple, s is called the subject term, p the predicate term and o the object term of that particular triple. A set of triples forms a graph G. Based on the positionality of terms appearing in the triples of G we distinguish between the subjects (S_G), predicates (P_G) and objects (O_G) of a graph. The nodes of a graph are defined as $N_G = S_G \cup O_G$.

Equivalence An equivalence relation $\equiv$ is a binary relation that is reflexive, symmetric and transitive. The identity relation is the smallest equivalence relation. The equivalence class of an RDF node $x \in N_G$ under $\equiv$ is $[x]_\equiv = \{y \in N_G \mid x \equiv y\}$.

Set theory We use the phrase "universe of discourse" to denote the instances that are formally described in a given dataset. Specifically, the universe of discourse for an

RDF graph G is N_G. We use the capital letters X and Y to denote arbitrary sets. Elements of these sets are denoted by $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ respectively.

Modeling identity It is common modeling practice to denote identity on the instance level with `owl:sameAs` and equivalence on the schema level with `owl:equivalentProperty` for properties and `owl:equivalentClass` for classes. We use $\sim$ to indicate a set of pairs that are explicitly specified to be the same, using either of these three properties. In addition, `owl:differentFrom` is used by modelers to indicate that two terms do not denote the same resource. We use $\nsim$ for a set of pairs that are explicitly indicated to *not* be the same.

Rough Set Theory Relations are called 'attributes' in Rough Set Theory. They are functions that map to an arbitrary set of value labels. We only consider functions that map from binary input into the set of Boolean truth values, and therefore use the term 'predicates' to denote these functions. We recognize that extensions to multi-valued logics would require a richer set of value labels. Rough Set Theory has been related to Formal Concept Analysis, e.g. in [8].

Formal Concept Analysis Formal Concept Analysis (FCA) takes a context $\langle O, A, M \rangle$ consisting of a set of objects O, a set of attributes A and a mapping M from the former to the latter. For a given set of objects $X \subseteq O$ one can calculate the attributes that are shared by those objects as $X' = \{y \in A \mid (\forall\, x \in X)(M(x, y))\}$. For a given set of attributes $Y \subseteq A$ one can calculate the objects that have (at least) those attributes as $Y' = \{x \in O \mid (\forall\, y \in Y)(M(x, y))\}$. A formal concept is a pair $\langle X, Y \rangle \in \mathcal{P}(O) \times \mathcal{P}(A)$ such that $X' = Y$ and $Y' = X$. The two functions $(\cdot)'$ are called the *polars* of M. For a given context, the set of concepts is denoted $\mathcal{B}(O, A, M)$. The concepts form a lattice $\langle \mathcal{B}(O, A, M), \{\langle\langle X_1, Y_1\rangle, \langle X_2, Y_2\rangle\rangle \in (\mathcal{P}(O) \times \mathcal{P}(A))^2 \mid X_1 \subseteq X_2\}\rangle$.

5 Approach

We start with a given identity relation $\sim$ that partitions the universe of discourse N_G into equivalence classes. Since the identity relation is the smallest equivalence relation, it is also the most fine-grained partition of N_G. As we saw in Principles 1 and 2, identity is indiscernibility with respect to all (possible) properties $\varPsi$. Besides identity, there are many other instances of indiscernibility: one corresponding to each set of properties $\varPhi \subseteq \varPsi$. According to this generalization, x and y are indiscernible with respect to a set of properties $\varPhi$ iff $(\forall\, \phi \in \varPhi)(\phi(x) = \phi(y))$. Every indiscernibility relation is also an equivalence relation, although not necessarily the smallest one. Every indiscernibility relation defined over domain N_G is also an identity relation, just over a different domain [19]. For instance, the set of properties $\varPhi = \{$"has an income of 1,000 euro's"$\}$ does not uniquely identify people (since two people may have the same income), but does uniquely identify income groups.

Let us consider two medicines Baspirin (`abox:baspirin`) and Caspirin (`abox:caspirin`) that both contain acetylsalicylic acid as their chemical compound (`tbox:chemComp`).

A chemist observes that they have the same substance and asserts that they are identical (`owl:sameAs` or $\sim$), resulting in the graph in Figure 1. However, Basperin and Casperin are produced (`tbox:prod`) by different companties: B Inc. (`abox:binc`) and C Inc. (`abox:cinc`). Basperin and Casperin cannot be told apart in a language that only contains the properties "is a" and "has chemical compound". However, if the language also includes the property "is produced by" then these medicines can be told apart. In other words: *we can look at the set of properties as a parameter that can be adjusted in order to obtain an equivalence relation that is more or less fine-grained, as required in different contexts* (in our example: contexts where the commercial supplier does or does not play a role in distinguishing two drugs).

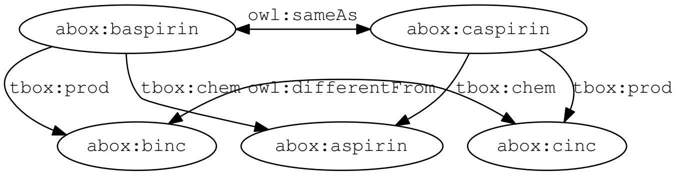

Fig. 1: Graph showing some of the assertions that we use as examples.

We now reinterpret the identity relation $\sim$ as if it were an indiscernibility relation $\approx_{\Phi}$ whose set of properties Φ is implicit in the data. Based on the extensional specification of the identity relation we can explicate the set of properties to which it is indiscernible with Definition 2, where $\{x_1, \ldots, x_n\}$ is one of the equivalence classes closed under $\sim$.

Definition 2 (Indiscernibility properties).

$$\begin{gathered} P^{+}(\{x_1, \ldots, x_n\}) = \{p \in P_G \mid (\exists\, p_1, \ldots, p_n \in [p]_{\equiv})(\\ [\{o \in O_G \mid \langle x_1, p_1, o\rangle\}]_{\equiv} = \ldots = [\{o \in O_G \mid \langle x_n, p_n, o\rangle\}]_{\equiv})\} \end{gathered}$$

For instance, by using Definition 2 we can deduce that the indiscernibility properties of Basperin and Casperin include `rdf:type`, `tbox:chemComp` and, by definition `owl:sameAs`. Notice that both the predicate and object terms are closed under identity. Performing these closures is important in order to identify the relevant indiscernibility properties. For instance, chemical compound, or `tbox:chemComp`, in one dataset may be the same property as chemical substance, or `ex:chemSubst`, in another. Besides the indiscernibility properties, there may also be discernibility properties (Definition 3), i.e., properties that indicate that two terms should *not* be considered to denote the same resource. As with the identity relation $\sim$, we assume that we are given a 'different-from' relation $\nsim$ of pairs $\langle x_1, x_2\rangle$.

Definition 3 (Discernibility properties).

$$\begin{aligned} P^{-}(\{x_1,x_2\}) = \{p \in P_G \mid (\exists\, p_1, p_2 \in [p]_{\equiv})(\\ \langle x_1,p_1,o_1\rangle, \langle x_2,p_2,o_2\rangle \in G \wedge (\exists \langle y_1,y_2\rangle \in\nsim)(o_1 \in [y_1]_{\equiv} \wedge o_2 \in [y_2]_{\equiv}))\}\end{aligned}$$

In our example, the discernibility properties for Basperin and Casperin includes `tbox:prod`. Using the indiscernibility and discernibility properties we can define the indiscernibility relation (Definition 4).

Definition 4 (Indiscernibility relation).

$$x \approx_{\Phi} y \;\Leftrightarrow\; P^{+}(\{x,y\}) = \Phi \wedge P^{-}(\{x,y\}) \cap \Phi = \varnothing$$

For our example we derive that Baspirin and Caspirin are the same with respect to the type and chemical compound properties (Example ex1) and that they are not the same with respect to the producer property (Example ex2). Another way of phrasing this is: Basperin and Caspirin are the same drug in terms of their chemical compound, but they are different medical products.

$$\texttt{abox:baspirin} \approx_{\{\texttt{owl:sameAs},\texttt{rdf:type},\texttt{tbox:chemComp}\}} \texttt{abox:caspirin} \tag{ex1}$$

$$\texttt{abox:baspirin} \not\approx_{\{\texttt{tbox:prod}\}} \texttt{abox:caspirin} \tag{ex2}$$

Now that we have defined the indiscernibility properties for a given set of resources, we go on to say that two pairs of resources are *semi-discernible* iff their indiscernibility properties are the same. When we look at the pairs that constitute (the extension of) a given identity relation $\sim$, all identity assertions look the same. But when we redefine identity in terms of indiscernibility and semi-discernibility, we see that within a given identity relation there are pairs that are indiscernible with respect to different properties. Stating this formally, semi-discernibility is an equivalence relation on pairs of resources that induces a partition of the Cartesian product of the universe of discource. Definition 5 makes this concrete in terms of the earlier definitions.

Definition 5 (Semi-discernibility).

$$\begin{aligned} \langle x_1,y_1\rangle \equiv_{\Phi} \langle x_2,y_2\rangle \;\Leftrightarrow\; & P^{+}(\{x_1,y_1\}) = P^{+}(\{x_2,y_2\}) = \Phi \\ & \wedge P^{-}(\{x_1,y_1\}) \cap P^{-}(\{x_2,y_2\}) = \varnothing \end{aligned}$$

For example, Baspirin and Caspirin are semi-discernible to Bicotine and Nicotine, two stimulant drugs (indiscernibility property `rdf:type`) whose chemical compound (indiscernibility property `tbox:chemComp`) is nicotine. An example of semi-discernible pairs from another application domain are $\langle$`dbr:Amsterdam`, `dbr:Rotterdam`$\rangle$ and $\langle$`dbr:Netherlands`, `dbr:Germany`$\rangle$, since the former are both cities and the latter are both countries (discernibility property `rdf:type` and each pair is part of the same geographic region (Amsterdam and Rotterdam are part of the Netherlands; the Netherlands and Germany are part of Europe).

Notice that the partitions obtained by $\equiv_{\Phi}$ contain but are not limited to the original identity pairs. Therefore, for sets of pairs closed under semi-discernibility we can distinguish between the following three categories:

1. All pairs in the set are identity pairs. This characterizes a consistent subrelation of the identity relation, since no semi-discernible pair is left out.
2. Only some pairs in the set are identity pairs. This characterizes a subrelation of the identity relation that is not applied consistently with respect to the semi-discernibility relation that can be observed in the data.
3. No pairs in the set are identity pairs. This characterizes a subrelation of the collection of pairs that is consistently kept out of the identity relation.

Each member of the semi-discernibility partition that is not of the third kind, i.e., every set of pairs that contains at least some identity pair, can be thought of as an identity subrelation. Not only is the uniform set of `owl:sameAs` assertions partitioned into subrelations, but each subrelation is described in meaningful terms that are drawn from the dataset vocabulary.

Now that we have determined the subrelations of identity we go on to define how these subrelations are related. Borrowing insights from Formal Concept Analysis we take N_G^2 as our set of FCA objects and P_G as our set of FCA attributes. The mapping from the former to the latter is $M(\langle x, y\rangle) = \Phi(\{x, y\})$. Because the number of FCA objects is quadratic in the size of the universe of discourse it is not practical to calculate the full concept lattice. However, we are only interested in the identity subrelations and how they are related to one another. Indeed, for every pair $\langle x, y\rangle \in\sim$ we can calculate the formal concept $\langle\{\langle x, y\rangle\}'', \{\langle x, y\rangle\}'\rangle$ by using the polars $(\cdot)'$. What FCA adds to the picture is a partial order $\leq$ between the identity subrelations (Definition 6).

Definition 6 (Indiscernibility lattice). *For a given identity relation $\sim$, the poset of indiscernibility subrelations is $\langle B, \leq\rangle$ with $B = \{\langle\{\langle x, y\rangle\}'', \{\langle x, y\rangle\}'\rangle \mid \langle x, y\rangle \in\sim\}$ and $\langle\{\langle x_1, y_1\rangle\}'', \{\langle x_1, y_1\rangle\}'\rangle \ \leq\ \langle\{\langle x_2, y_2\rangle\}'', \{\langle x_2, y_2\rangle\}'\rangle$ iff $\Phi(\{x_1, y_1\}) \subseteq \Phi(\{x_2, y_2\})$.*

Every node in the lattice corresponds to a different set of indiscernibility properties, i.e., to a different subrelation of the identity relation. Each subrelation corresponds to an identity assertion context. Specifically, the indiscernibility properties Φ denote the aspects that are important in that context. Results derived/entailed in one context may not be derived in another. Asserting/retracting statements changes the indiscernibility lattice (even if the identity relation is kept the same). The indiscernibility lattice for the graph in Figure 1 is given in Figure 2.

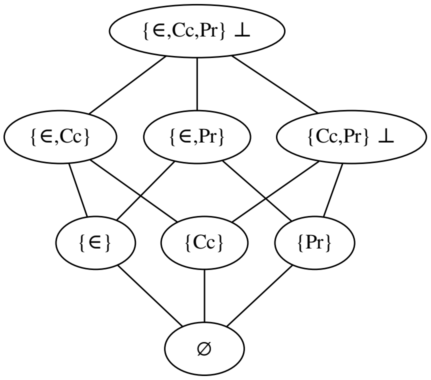

Fig. 2: The indiscernibility lattice for the graph in Figure 1. For readability we abbreviate `tbox:chemComp` as Cc, `tbox:prod` as Pr, `rdf:type` as $\in$, and `owl:sameAs` as $\sim$. Two of the indiscernibility relations cannot be chosen without resulting in an inconsistent state.

Now that we have defined the indiscernibility lattice that comes with a given identity relation $\sim$, we can define the possible identity contexts (Definition 7) in which only part of an identity relation can be used, namely the part that is relevant relative to that identity context.

Definition 7 (Identity context). *For a given identity relation $\sim$ and its indiscernibility lattice $\langle B, \leq \rangle$ an identity context is a subset of formal concepts $B' \subseteq B$ such that for all $\langle o_1, a_1 \rangle, \langle o_2, a_2 \rangle \in B'$ we have that (i) $a_1 \cap a_2 = \varnothing$ and (ii) $(\forall x, y \in N_G)(P^+(\{x, y\}) \nsubseteq a_1 \vee P^-(\{x, y\}) \nsubseteq a_2)$.*

6 Applications

Modeling In Figure 2 every node denotes an indiscernibility relation $\approx_\Phi$ based on a different set of indiscernibility properties Φ. We can define the *precision* of each node by quantifying how many of the pairs that are indiscernible with respect to Φ are also in the original identity relation: $|\sim \cap \approx_\Phi| \;/\; |\approx_\Phi|$. We can also define the *recall* of each node by quantifying how much of the original identity relation is characterized by Φ: $|\sim \cap \approx_\Phi| \;/\; |\sim|$.

The identity lattice annotated with precision and recall numbers can be used to deliver feedback to the modeler. For instance, low precision nodes indicate the absence of identity criteria that are explicit in the data. In practice, many identity links depend on special knowledge the modeler had at the time of assertion. If such special knowledge is not encoded in the data then another data user can no longer validate whether these links are correct. Automatic calculation of the precision of nodes in the identity lattice may prompt a modeler to either (i) make the identity criteria explicit or (ii) remove the identity assertion altogether. The latter may be the case for very low precision nodes, possibly indicating accidental or erroneous identity assertions. Another way in which the identity lattice can support the modeler is by using high-precision nodes in order to give automated suggestions for identity assertion. Specifically, pairs that are indiscernible according to the same criteria as many of the identity pairs may be considered good candicates for identity assertion.

Reasoning with inconsistencies As we saw in Section 3, one of the main problems of the current use of identity is that terms are considered the same in some but not all contexts. As we saw in Section 2 this either results in too many entailments and contradictions, or it results in the use of syntactic alternatives like `skos:related` that do away with entailment altogether. An example of the former can be given with respect to the example shown in Figure 1, where the identity assertion of the two medicines based on their shared chemical compound results in the substitution of the two medicines in other contexts as well. Specifically, following the OWL2 rule in ent1 we derive that both medicines are produced by companies B Inc. and C Inc., which is unlikely to be the case.

$$\langle s, p, o \rangle \land s \sim s' \Rightarrow \langle s', p, o \rangle \tag{ent1}$$

Now that we have the indiscernibility lattice from Figure 2 we can choose an identity context that can be used to calculate some, but not all entailments. This is supported by condition (ii) in Definition 7 that excludes contexts that result in an inconsist state. The OWL2 rule in ent1 is adapted to take into account an identity context Con, resulting in rule ent2. Other entailment rules require similar adaptations.

$$(\exists \varPhi \in Con)(\langle s, p, o \rangle \land s \approx_{\varPhi} s' \land p \in \varPhi \Rightarrow \langle s', p, o \rangle \tag{ent2}$$

Quality assessment Borrowing insights from Rough Set Theory we can determine the quality of a given identity relation. The lower approximation of identity is the union of the indiscernibility relations that only contain identical pairs (Definition 8a). The higher approximation of identity is the union of indiscernibility relations that contain some identical pair (Definition 8b).

Definition 8 (Lower and higher approximation).

$$x_1 \underline{\sim} y_1 \;\Leftrightarrow\; \forall_{\langle x_2, y_2\rangle \in N_G^2}(\langle x_1, y_1\rangle \equiv_{\varPhi} \langle x_2, y_2\rangle \rightarrow x_2 \sim y_2) \tag{8a}$$

$$x_1 \overline{\sim} y_1 \;\Leftrightarrow\; \exists_{\langle x_2, y_2\rangle \in N_G^2}(\langle x_1, y_1\rangle \equiv_{\varPhi} \langle x_2, y_2\rangle \wedge x_2 \sim y_2) \tag{8b}$$

Based on these two approximations we can give the rough set representation $\langle \underline{\sim}, \overline{\sim} \rangle$ of the identity relation $\sim$ [18]. The quality of a rough set representation is given in Definition 9 and is always a number in $[0, 1]$.

Definition 9 (Quality). $\alpha(\sim) \;=\; |\underline{\sim}| \;/\; |\overline{\sim}|$

The quality of the identity relation is higher if the two approximations are closer to each other, and quality is highest if the two approximations are the same. The intuition behind this is that in a high-quality dataset the identity relation should be based on indiscernibility criteria that are explicit in the data. Formally this means that the semi-discernibility partition should consist of partition members that contain either no identity pairs (small value for $\overline{\sim}$) or only identity pairs (large value for $\underline{\sim}$). If a member of the semi-discernibility partition contains only some identity pairs then this means that the difference between identical and non-identical pairs cannot be based on the properties that are asserted in the data. As with the per-node precision and recall calculations (Section 6), the use of data-external identity criteria makes it more difficult to validate identity statements. The quality of a dataset can be improved by making explicit the properties two entities must share in order for them to be considered the same. Adding such indiscernibility properties results in a higher quality metric.

7 Implementation

The approach outlined in Section 5 was implemented and tested on datasets published in the instance matching track of the Ontology Alignment Evaluation Initative. Figure 3 shows an indicative example of an indiscernibility lattice that is calculated for such datasets. Each rectangular box represents an indiscernibility relation. The set notation shows the indiscernibility properties $\varPhi$ for each indiscernibility relation. For each box the precision quantifies how many pairs that are indiscernible with respect to $\varPhi$ are in the original identity relation, i.e., $|\sim \cap \approx_{\varPhi}| \;/\; |\approx_{\varPhi}|$. For each box the recall quantifies how much of the original identity relation is characterized by $\varPhi$, i.e., $|\sim \cap \approx_{\varPhi}| \;/\; |\sim|$.

Since in this Figure a partition is only drawn when there is at least one identity pair that is indiscernible with respect to some set of predicates, the higher approximation amounts to the entire figure. The lower approximation only consists of those partition sets that contain at least one identity pair, and that contain no non-identity pair; these are distinguished by green borders. For each box the precision number indicates the

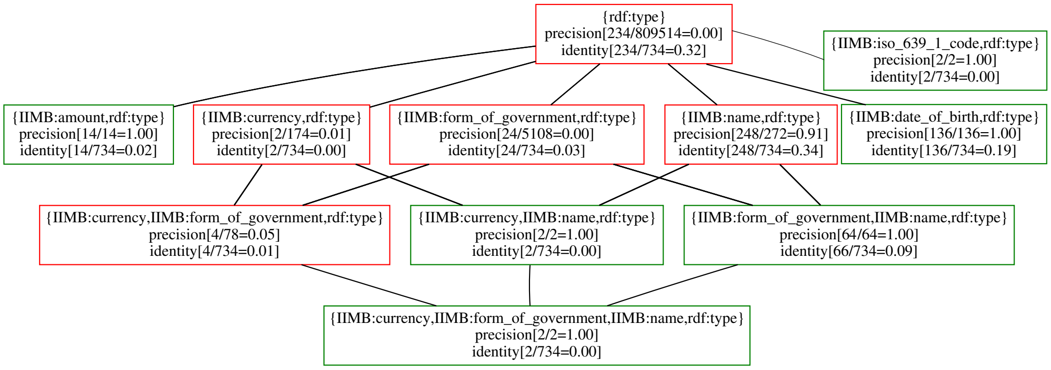

Fig. 3: Example of the identity subrelations for a dataset in the instance matching track of the Ontology Alignment Evaluation Initiative. This is the 16^{th} variant of the IIMB datasets in the 2012 challenge.

ratio of identity pairs for each subrelation. By definition, subrelations in the lower approximation have precision 1.0 and that subrelations in the higher approximation have a non-zero precision.

Figure 3 shows that the uniform identity relation consists of conceptually different indiscernibility subrelations. For instance, some entities are considered the same based on their $\{$`IIMB:amount`, `rdf:type`$\}$ properties (movies with the same budget are indiscernible in this dataset) and some entities are considered the same based on their $\{$`IIMB:date_of_birth`, `rdf:type`$\}$ properties (people with the same birth data are indiscernible in this dataset). Notice that in both cases strict identity would indeed be too strong, since two movies might have the same budget and two people might have the same birth data. The figure also shows that approximately 30% of the given identity relation extension is applied consistently with respect to the calculated indiscernibility lattice, i.e., the green boxes. The red boxes with high precision are able to isolate a limited number of pairs that are indiscernible in the same way as identity pairs but that are not in the given identity relation. An example of this is $\{$`IIMB:name`, `rdf:type`$\}$. These may either be candidates for identity assertions under the same condition, or some additional facts may be asserted about them in order to distinguish them from identity pairs. Finally, the figure shows that approximately one third of the original identity relation's extension are only indiscernible with respect to their `rdf:type` property. This is insufficient to set them apart from many non-identity pairs and results in a lower quality metric.

Calculation of the indiscernibility lattice is implemented in SWI-Prolog and its ClioPatria triple store [22]. Identity statements are either loaded from VoID linksets or are loaded from EDOAL (Expressive and Declarative Ontology Alignment Language) alignment files. The code is available at `http://github.com/wouterbeek/IOTW/`. For the 60 IIMB datasets in the OAEI 2012 Instance Matching track this naive implementation on average takes 15 seconds to calculate the identity lattice.

8 Conclusion

The identity relation `owl:sameAs` is a crucial element of the Semantic Web. It is therefore alarming that its semantics is both computationally ineffective and epistemilogically inadequate. Computationally, it is in principle impossible to define an effective procedure for establishing the truth of `owl:sameAs` assertions, because the open world assumption implies that the set of properties to be checked for indiscernibility is unknown; and epistemilogically it is impossible to model the situation that two given objects may be regarded as equal in one context, but not equal in another.

In this paper we presented a new approach for defining the identity relation. Instead of checking indiscernability with respect to all properties we explicity parameterise the identity relation over the set of properties that are taken into account for establishing identity. This gives both a computationally effective procedure and allows us to define different identity relations in different contexts.

Section 3 enumerates four desiderata for a semantics of identity. (i) The semi-discernibility partition allows the uniform identity relation to be characterized in terms of discernibility subrelations based on different sets of properties Φ. (ii) Since entailment can be defined with respect to a context, or collection of discernibility properties, it can be scoped to contexts in which entities are considered identical, preserving some of the benefits of entailment without resulting in an inconsistent state. (iii) Since the criteria for the discernibility subrelations are explicit in the data, the new semantics opens up possibilities for providing feedback to the modeler. (iv) A quality metric can be calculated for the identity relation of a dataset, indicating the consistency with which identity can be described in terms of the properties that occur in the data, rather than being based on knowledge left implicit by the original modeler.

The implications for OWL2 entailment under the here proposed semantics must be further investigated. Existing entailment languages such as RIF must be extended so that an identity context can be expressed. The current implementation is only a naive proof of concept and needs to be improved by using recent advances in calculating FCA's, e.g. [21], in order to be applicable to larger datasets. Quality metrics for identity could extend existing data quality metrics. Finally, the feedback mechanisms that are supported by the here presented semantics may be implemented as a plugin for an often used modeling editor such as Protégé in order to allow the utility of such features to be measured in practice.

References

1. Keith Alexander, Richard Cyganiak, Michael Hausenbals, and Jun Zhao. Describing linked datasets with the VoID vocabulary, March 2011.
2. Grigoris Antoniou, Paul Groth, Frank van Harmelen, and Rinke Hoekstra. *A Semantic Web Primer (Third Edition)*. The MIT Press, 2012.
3. Tim Berners-Lee. Linked Data. `http://www.w3.org/DesignIssues/LinkedData.html`, 2010.
4. Gerard de Melo. Not quite the same: Identity constraints for the web of linked data. In *Proceedings of the American Association for Artificial Intelligence 2013*, 2013.
5. Li Ding, Joshua Shinavier, Tim Finin, and Deborah L. McGuinness. OWL:sameAs and Linked Data: An empirical study. In *Proc. of the WebSci*, 2010.
6. Li Ding, Joshua Shinavier, Zhenning Shangguan, and McGuinness Deborah. SameAs networks and beyond: Analyzing deployment status and implications of owl:sameAs in Linked Data. In *The Semantic Web – ISWC 2010*, volume 6496, pages 145–160. 2010.
7. Peter Forrest. The identity of indiscernibles. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. 2008.
8. Bernhard Ganter. Non-symmetric indiscernibility. In *Knowledge Processing and Data Analysis*, pages 26–34. Springer, 2011.
9. P.T. Geach. Identity. *Review of Metaphysics*, 21:3–12, 1967. Reprinted in Geach 1972, pp. 238–247.
10. Harry Halpin, Patrick Hayes, James McCusker, Deborah McGuinness, and Henry Thompson. When owl:sameas isn't the same: An analysis of identity in linked data. In *The Semantic Web – ISWC 2010*, volume 6496, pages 305–320. 2010.
11. Harry Halpin, Patrick J Hayes, and Henry S Thompson. When owl:sameas isn't the same redux: Towards a theory of identity, context, and inference on the semantic web. In *Modeling and Using Context*, pages 47–60. Springer International Publishing, 2015.
12. Saul Kripke. *Naming and Necessity*. Cambridge: Harvard University Press, 1980.
13. David Lewis. *On the plurality of worlds*. Oxford: Basil Blackwell, 1986.
14. Chang Liu, Guilin Qi, Haofen Wang, and Yong Yu. Fuzzy reasoning over RDF data using OWL vocabulary. In *Proceedings of the Int. Conf. on Web Intell. and Intell. Agent Technology*, pages 162–169, 2011.
15. James McCusker and Deborah McGuinness. Towards identity in linked data. *Proceedings of OWL Experiences and Directions Seventh Annual Workshop*, 2010.
16. Alistair Miles and Sean Bechhofer. SKOS simple knowledge organization system reference, August 2009.
17. Boris Motik, Bernardo Cuenca Grau, and Peter Patel-Schneider. OWL 2 web ontology language direct semantics (second edition), December 2012.
18. Zdzisław Pawlak. *Rough Sets: Theoretical Aspects of Reasoning About Data*. Kluwer Academic Publishing Dordrecht, 1991.
19. Willard Van Orman Quine. Identity, ostension, and hypostasis. *The Journal of Philosophy*, 47(22):621–633, 1950.
20. Tommaso Soru, Edgard Marx, and Axel-Cyrille Ngonga Ngomo. Rocker: A refinement operator for key discovery. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1025–1033. International World Wide Web Conferences Steering Committee, 2015.
21. Vychodil Vilem. A new algorithm for computing formal concepts. *Cybernetics and Systems*, pages 15–21, 2008.
22. Jan Wielemaker, Wouter Beek, Michiel Hildebrand, and Jacco van Ossenbruggen. Cliopatria: A logical programming infrastructure for the semantic web. *Semantic Web Journal*, 2015.