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 the Linked Open Data initiative and on the Semantic Web in general. They allow the interlinking of alternative descriptions of the same thing. However, many practical uses of `owl:sameAs` are known to violate its formal semantics. We propose a method that assigns meaning to (the subrelations of) an identity relation using the predicates of the dataset schema. Applications of this approach include automated suggestions for asserting/retracting identity pairs and quality assessment. We also describe an experimental design for this approach.

Introduction

Identity relations are at the foundation of the Linked Open Data initiative and of the Semantic Web in general (Bizer, Cyganiak, and Heath 2007). They allow the interlinking of alternative descriptions of the same thing. However, the traditional notion of identity (expressed by `owl:sameAs` (W3C 2012)) is often problematic, e.g. when objects are considered the same in some contexts but not in others. The standing practice in such cases is to use weaker relations of relatedness (e.g., `skos:related`). Unfortunately, this limits reasoners in drawing inferences.

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. 2010b; 2010a).

Previous work

Existing research proposes the following solutions for the problem of identity relations on the Semantic Web. (1) Introduce weaker versions of `owl:sameAs` (Halpin et al. 2010b; McCusker and McGuinness 2010) (e.g., `skos:related`). (2) Restrict the applicability of identity relations to specific contexts. Identities are expected to hold within a named graph or within a namespace, but not necessarily outside of it (Halpin et al. 2010b; de Melo 2013). (3) Introduce additional vocabulary that does not weaken but extend the existing identity relation (Halpin et al. 2010b). For example, allow an explicit distinction to be made between mentioning a term and using a term (e.g., a car and a Web document describing that car). (4) Add domain-specific weaker versions of the identity relation (McCusker and McGuinness 2010) (e.g., “have the same medical use” is weaker than “are the same molecule”). (5) Adapt the 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 respective datasets. (Halpin et al. 2010b; Ding et al. 2010a)

Other related research focusses on the extraction of network properties of `owl:sameAs` datasets (Ding et al. 2010b), but these endeavors are not yet related to the semantics of the identity relation.

What these approaches have is common is that quite some work has to be done (adapting or creating standards, instructing modelers, converting existing dataset) in order to resolve 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 is fully automated and that requires no changes to standards, modeling practices, or existing datasets.

Research goals

In developing our approach 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 occur in the schema of the dataset.
2. Based on an existing identity relation we want to give semantically motivated suggestions for extending/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.

Approach

In the following we consider an arbitrary, materialized RDF graph G and an identity relation $\sim$ over the resources that occur in G. The subject terms of G are denoted by S_G, the predicate terms by P_G and the object terms by O_G.

For illustrative purposes we use the IIMB dataset that is used in the instance matching track of the 2012 Ontology Alignment Evaluation Initiative (OAEI) (OAEI 2012). This dataset consists of eighty ontologies G_i (for $1 \leq i \leq 80$) that are linked to a single base ontology G_0. A graph G is the result of fully materializing the graph merge of G_i (for some $1 \leq i \leq 80$) and G_0. For each of these eighty linked ontologies a reference mapping is available.

In RDF a property consists of a single predicate term (e.g., the property "is spoken in" is denoted by predicate term `IIMBTBOX:spoken_in` in the IIMB dataset). We generalize the notion of a property to consist of an arbitrary number of predicate terms. For this we define a depth-n *predicate path map* (abbr. ppm) that is characterized by a sequence of n predicates and denotes a (functional) mapping from subject terms into sets of object terms (def. 1, where $p_i \in P_G$ and $[p]_\sim$ is the identity set for p).

Definition 1 (Predicate path map (ppm))**.**

$$f_{\langle p_1,\ldots,p_n\rangle}(s) = \{o \in O_G \mid \exists_{x_1,\ldots,x_{n-1}} \bigwedge_{i=0}^{n-1} \langle I(x_i), I(x_{i-1})\rangle \in \bigcup_{p\in[p_{i+1}]_\sim} Ext(I(p))\} \tag{1}$$

Examples of properties characterized by predicate path maps are "is spoken in" (depth-1) and "is spoken in a country whose form of government is" (depth-2).

Shared properties and indiscernibility

When we look at the triples that constitute a set of identity relations, we see that all links look the same. But when we take the triples in which the subject and object terms occur into account, we see that within the identity relation there may be different subrelations that we can identify in terms of the predicates that occur in the schema.

For instance, in the IIMB dataset there are some identical resources that share the property `IIMBTBOX:spoken_in`, while other pairs share the property `IIMBTBOX:form_of_government`. The set of pairs of resources that are spoken in the same language may even be disjoint from the set of pairs of resources that have the same form of government.

Note that we are not only interested in the properties that resources share with one other (e.g., where they are spoken, or which form of government they have), but we are also interested in resource pairs that share the same sharing properties. We can thus identify subsets of an identity relation based on differences in the sets of predicate path maps relative to which they take resources to be *indiscernible* from one another.

In the example above, one subset of the identity relation does not discern resources that are spoken in the same language, whereas another subset of the identity relation does not discern resources that have the same form of government.

We say that two resources are indiscernible with respect to a set of predicate path maps $P \subset P_G^n$ in case they share the same properties denoted by those ppms (def. 2).[1] We say that two resource pairs are indiscernible in case both pairs are indiscernible for the same $P^* \subseteq \mathcal{P}(P_G^n)$ (def. 3).[2]

Definition 2 (Indiscernibility)**.**

$$IND(P) = \{\langle x,y\rangle \in S_G^2 \mid \forall_{p\in cl_\sim(P)} f_p(x) \approx f_p(y)\} \tag{2}$$

$$IND(P^*) = \{\langle\langle x_1,y_1\rangle,\langle x_2,y_2\rangle\rangle \in (S_G^2)^2 \mid \forall_{P\in P^*} \langle x_1,y_1\rangle \in IND(P) \leftrightarrow \langle x_2,y_2\rangle \in IND(P)\} \tag{3}$$

As explained above, for a given set of identity pairs there may be multiple pairs that have the same shared properties. These sets of predicates that are shared across resource pairs are considered to give a description of a specific subrelation of the identity relation.

According to the standard definition, identical resources are indiscernible with respect to all properties. We take a given set of identity pairs and partition it into subsets which we can describe as being $cl_\sim(P)$-indiscernible, for $P \subseteq P_G^n$.

Fig. 1 shows an example of a discernibility partitioning for a given identity relation.

Approximation

In the previous section we partitioned a given identity relation $\sim$ into subrelations that can be distinguished in terms of schema predicates (or n-depth paths of those predicates). In this section we create an approximation of the identity relation. This approximation will allow us to (1) give suggestions about which pairs to in/exclude from the identity relation, and (2) give an indicator for the quality of the identity relation.

For the approximation of the identity relation we use rough set theory (Pawlak 1991) to represent an approximation of a given identity relation $\sim$. The domain for our rough set approach is the Cartesian product of S_G. The set of predicates is the powerset of P_G.

For an arbitrary binary relation $\sim$ we can define a higher (def. 4) and a lower (def. 5) approximation of that relation. In definitions 4 and 5, $\mathbb{R}$ characterizes a similarity relation between resource pairs. The intuition behind these definitions is that non-$\sim$-pairs that are similar to $\sim$-pairs should be in the higher approximation, whereas no $\sim$-pair that has a similar non-$\sim$-pair should be in the lower approximation.

Definition 3 (Higher & lower approximation)**.**

$$x \overline{\sim} y \iff \exists u,v(\langle u,v\rangle \mathbb{R} \langle x,y\rangle \land u \sim v) \tag{4}$$

$$x \underline{\sim} y \iff \forall u,v(\langle u,v\rangle \mathbb{R} \langle x,y\rangle \rightarrow u \sim v) \tag{5}$$

Since we want to stay close to the traditional notion of identity, defined in terms of indiscernibility, we choose $IND(\mathcal{P}(P_G^n))$ as our similarity relation.

Figure 1 shows an example of the lower and higher approximations for a linkset. Since in this figure a partition is

[1] P must be closed under the identity relation, i.e.,

$$cl_\sim(P) = \bigcup_{\langle p_1,\ldots,p_n\rangle\in P} ([p_1]_\sim \times \ldots \times [p_n]_\sim)$$

[2] In order to assertain that $f_p(x)$ and $f_p(y)$ denote the same set of resources, identity does not suffice. $\approx$ gives a special treatment for blank nodes and typed literals (skipped here for brevity).

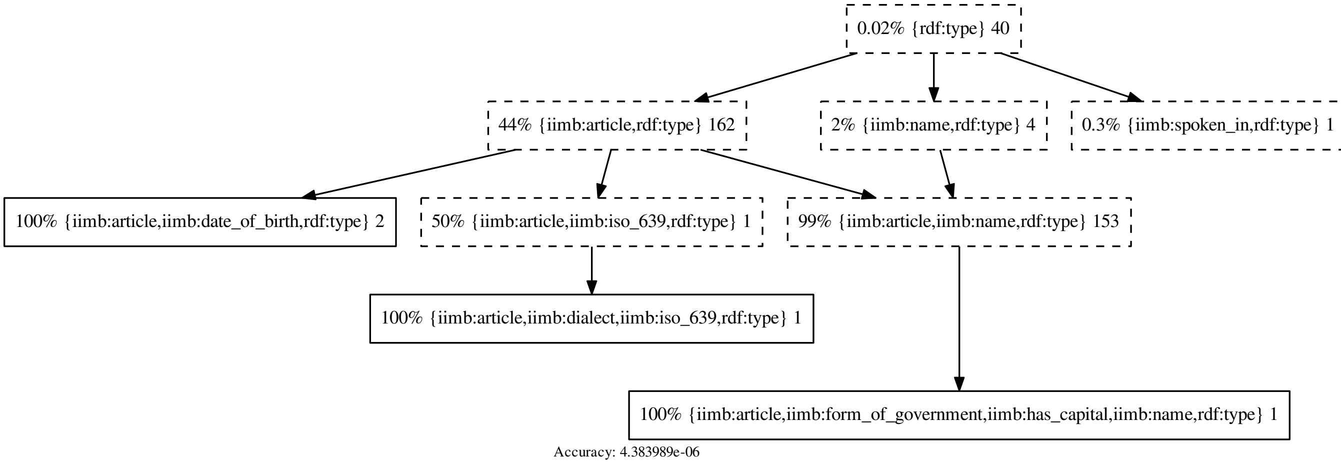

Figure 1: An example of a discernibility partition for an identity relation consisting of 365 pairs applied to the fourth IIMB linkset. Each node is annotated with the set of predicates P for which its pairs are P-indiscernible. The number of identity pairs within each partition set is displayed to the right of the predicate set label. Partition sets that contain no identity pair are not show. The number that occurs to the left of the predicate label in each node indicates how may pairs in that node are identity pairs. The lower approximation consists of the nodes with a solid border, indicating that they contain only identity pairs. The higher approximation consists of all displayed nodes.

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.

Quality

Given the rough set representation $\langle \underline{\sim}, \overline{\sim} \rangle$ of identity relation $\sim$, we can calculate the accuracy of this approximation with equation 6.

$$\alpha(\sim) = |\underline{\sim}|/|\overline{\sim}| \tag{6}$$

The intuition behind the usefulness of equation 6 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 $\overline{\sim}$) or only identity pairs (big value for $\underline{\sim}$), a more consistent identity relation has a higher accuracy.

Now that we have a formal metric for identity relation quality, we can define the characteristics of an ideal identity relation. Traditionally the ideal identity relation ensures indiscernibility for all expressible properties in the language (the principle of the indiscernibility of identicals). According to this traditional view an identity relation becomes of higher quality by considering more predicates (or ppms) according to which two resources are not allowed to be discernible. We give a different quality criterion.

We observe that for a given equivalence relation $\sim$ defined over a domain of resources S_G we can define the notion of full discernibility:

Definition 4 (Discernible model)**.**

A domain S_G is fully discernible w.r.t. a binary relation $\sim$ iff

$$\forall x, y \in S_G([x]_\sim = [y]_\sim \lor \mathbb{P}([x]_\sim) \neq \mathbb{P}([y]_\sim)) \tag{7}$$

From this definition it is clear that a domain of discourse is fully discernable just in case there exists a binary relation $\sim$ such that $\alpha(\sim) = 1.0$.

Experimental design

In section we enumerated three research goals. The first goal is met, since an indiscernibility partition characterizes subrelations based on the ppms P for which the pairs in that sets are $cl_\sim(P)$-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.

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 (i.e., “not must not”, 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.

Our approach provides a new experimental design for evaluating hypothesis regarding identity relations that have not been evaluated before in terms of the semantics of the data.

Hypotheses

Using this new approach the following hypothesis can be validated:

1. Take an `owl:sameAs` relation and a `skos:related` relation defined over the same domain. Merge them into a new binary relation $\sim$. Establishing the lower and higher approximation of $\sim$, the hypothesis is that pairs from `owl:sameAs` occur more frequently in the lower boundary than pairs from `skos:related`.
2. Take a set of alignment pairs, each of which is associated with a confidence measure between 0.0 and 1.0. Choose an arbitrary cutoff point $0.0 < c < 1.0$. The hypothesis is that alignments with a confidence larger than c occur more frequently in the lower approximation than alignments with a confidence smaller than c.
3. Take a set of automatically generated alignment pairs with associated confidence measures and take the gold standard or reference alignment for the same dataset. The hypothesis is that pairs that occur in the lower approximation of the alignment appear relatively more often in the gold standard than pairs that occur in the higher approximation of the alignment.
4. The accuracy measure α of a reference alignment is generally higher than the accuracy measure of an automatically generated alignment for the same dataset. Or, the accuracy measure is generally higher for identity relations that are considered correct by domain experts.

Implementation

The implementation built for this paper is deployed as an extension pack of the ClioPatria triple store (Schreiber et al. 2006). For RDF graphs that are loaded in ClioPatria this extension calculates the discernibility partition, rough set approximation and accuracy. The results are visualized using GraphViz and are displayed in a Web interface using SVG. Interactive Ajax code allows the user to click on nodes in the SVG graphic to navigate to descriptions of the resource pairs that occur in that partition set while not being in the identity relation. This implementation may facilitate the validation of hypotheses in this new experimental setup.

Conclusion

In this paper we have given an new approach for characterizing, extending/retracting, and assessing identity relations. Our approach does this in purely qualitative terms, using schema semantics. In contemporary ontology alignment and data linking activities nonsemantic aspects of resources play a role as well. For instance similarity assessment for natural language labels is often used in data linking.

We think that the qualitative means of characterizing an identity relation are a useful addition to existing quantitative means. Also, we think that it is more useful and viable to enrich existing identity relations in the LOD based on the semantics of the datasets in which they occur, than to introduce new relationships into SW languages. Apart from the practical difficulties of teaching practitioners and transforming/enriching existing datasets, we suggest that the meaning of an identity (sub)relation is partially defined in its use, i.e., in the indiscernibility criteria it embodies.

For our approach it is not necessary to pose additional restrictions on a binary relation $\sim$. The definitions in this paper apply to `owl:sameAs` relations in the same way in which they apply to any other binary relation (e.g., `skos:related`).

We are currently in the process of validating the above enumerated hypotheses. The results of these evaluations are continuously being published on `wouterbeek.com/identity-on-the-web`. The website currently contains the automated results of all eighty IIMB alignments, drawn from the instance matching track of the OAEI 2012. The website also refers to the publicly available Git repository `github.com/wouterbeek/IOTW` where the implementation discussed in section can be found.

References

Bizer, C.; Cyganiak, R.; and Heath, T. 2007. How to publish linked data on the web. Last accessed in May 2013.

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.; 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*. Lecture Notes in Computer Science. Springer Berlin / Heidelberg. 305–320.

Halpin, H.; Hayes, P.; McCusker, J.; McGuinness, D.; and Thompson, H. 2010b. When owl:sameas isnt 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.

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

OAEI. 2012. The instance matching track of the ontology alignment evaluation initative (oaei) 2012.

Pawlak, Z. 1991. *Rough Sets: Theoretical Aspects of Reasoning About Data*. Kluwer Academic Publishing Dordrecht.

Schreiber, G.; Amin, A.; van Assem, M.; de Boer, V.; Hardman, L.; Hildebrand, M.; Hollink, L.; Huang, Z.; van Kersen, J.; de Niet, M.; Omelayenko, B.; van Ossenbruggen, J.; Siebes, R.; Taekema, J.; Wielemaker, J.; and Wielinga, B. 2006. Multimedian e-culture demonstrator. *The Semnantic Web - ISWC 2006, Athens, Georgia* 4273:951–958. Winner of the Semantic Web Challenge 2006.

W3C. 2012. Owl 2 web ontology language direct semantics (second edition).