When owl:sameAs isn't the Same: An Analysis of Identity in Linked Data

Harry Halpin, Patrick J. Hayes, James P. McCusker, Deborah L. McGuinness, and Henry S. Thompson
{hhalpin,ht}@inf.ed.ac.uk, phayes@ihmc.us, {mccusj,dlm}@cs.rpi.edu

School of Informatics
University of Edinburgh
10 Crichton St.
EH8 9LW Edinburgh, UK
and
Institute for Human and Machine Cognition
40 South Alcaniz St.
Pensacola, FL 32502 USA
and
Tetherless World Constellation
Department of Computer Science
Rensselaer Polytechnic Institute
110 8th Street, Troy, NY 12180 USA

Abstract. In Linked Data, the use of *owl:sameAs* is ubiquitous in interlinking data-sets. There is however, ongoing discussion about its use, and potential misuse, particularly with regards to interactions with inference. In fact, *owl:sameAs* can be viewed as encoding only one point on a scale of similarity, one that is often too strong for many of its current uses. We describe how referentially opaque contexts that do not allow inference exist, and then outline some varieties of referentially-opaque alternatives to *owl:sameAs*. Finally, we report on an empirical experiment over randomly selected *owl:sameAs* statements from the Web of data. This theoretical apparatus and experiment shed light upon how *owl:sameAs* is being used (and misused) on the Web of data.

Keywords: *linked data, identity, coreference*

1 Introduction

As large numbers of independently developed data-sets have been introduced to the Web as Linked Data, the vexing problem of identity has returned with a vengeance to the Semantic Web. As the ubiquitous *owl:sameAs* property is used as the RDF property to connect these data-sets, it has been dubbed the '*owl:sameAs* problem' by publishers and users of Linked Data. However, the problem of identity lies not within Linked Data *per se*, but is a long-standing

and well-known issue in philosophy, the problem of identity and reference. What precisely *is* new in the recent appearance of this problem on the Web of Linked Data is that this is the first time the problem is being encountered by different individuals attempting to *independently* knit their knowledge representations together using the same standardized language. Much of the supposed "crisis" over the proliferation of *sameAs* in Linked Data can be traced to the fact that many mutually incompatible intuitions motivate the use of *owl:sameAs* in Linked Data. These intuitions almost always violate the rather strict logical semantics of identity demanded by *owl:sameAs* as officially defined.

To review, the *owl:sameAs* (abbreviated from hereon simply *sameAs*) construct is defined as stating "that two URI references actually refer to the same thing" [3]. For example, the city of Paris is referenced in a number of different Linked data-sets: ranging from OpenCyc to the New York Times. For example, we find that *dbpedia:Paris* is asserted to be *sameAs* both *cyc:CityOfParisFrance* and *cyc:Paris_DepartmentFrance* (and five other URIs). Yet OpenCyc explicitly states (in English!) that these two are distinct. Is there a contradiction here? Is DBPedia misusing *sameAs*? In this paper we will explore the origins of this (very common) situation, and suggest some ways forward.

As the Semantic Web is a project in development, it is always possible to specify anew various constructs. The project of inspecting alternative readings of *sameAs* has been begun by us in the past by looking at context [9] and proposed ontologies [12]. In this work we bring our research together and validate it empirically. We begin by reviewing the philosophical origin of the problem of identity from Leibnitz's Law in Section 2 and its implementation as *sameAs* in Section 3. In Section 4 we demonstrate a number of theoretically-motivated distinctions that are 'kind of close' to *sameAs* and then systematize these into an ontology in Section 5. Finally, test see if humans can reliably use these distinctions in Section 6, and conclude with recommendations for the future development of RDF in Section 7.

2 What is Identity?

The father of knowledge representation, Leibnitz, was also the first to phrase a coherent and formalizable definition of identity, often called 'Leibnitz's Law' or the 'The Identity of Indiscernables,' namely that that if x is not identical to y, then there must be some property that they do not share [11]. Or put another way, if x and y share all properties (i.e. if they are indiscernable) then they are identical. This law can then stated logically as $\forall x \forall y \exists P . x \neq y \rightarrow P(x) \wedge \neg P(y)$. The inverse of this is the more trivial law of substitutivity, which can then be stated as $\forall x \forall y . P(x) \wedge P(y) \rightarrow x = y$. Leibnitz's law and the law of substitutivity, which are obvious from a logical perspective, have a number of very practical engineering reprecussions in a distributed knowledge representation system such as the Semantic Web.

A number of classical problems already crop up in this analysis of identity. For example consider changes over time. Should things with different temporal-

spatial co-ordinates be counted as different, even if they share the rest of their properties? While that sounds like a common-sense distinction, is Tim Berners-Lee as an adult is the same as Tim Berners-Lee five minutes ago? Or as a child? Or if he lost his arm? This leads straight in to arguments about perdurance and endurance in philosophy. In any engineering discipline such as knowledge representation (as opposed to say, metaphysical thought experiments), we can *never* enumerate all possible properties.

Instead, we consider only a subset of possible properties. As a result identity based on propery matching is under-determined. One solution is to have *some* properties count as those necessary for identity, namely an explicit *theory of identity criteria*. Are there two different kinds of properties, properties that are somehow *intrinsic* to identity and others that are *extrinsic*, i.e. purely relative to other things?[1] However, this does not mean that all such criteria-based theories are compatible. One can imagine theories of identity based on different criteria, where some theories of identity subsume weaker or stronger ones, but others are simply incommensurable. Problems also arise with respect to (comparing) property *values*, for example when values are vague (is "purple" the same as "rgb(255,0,255)") or imprecise (is "2 inches" the same as "5 cm").

Regardless of these well-known issues, the point of a logical analysis of identity is clear in terms of inference: When someone says two things are the same, *the two things share all the same properties* and so every property of one thing can be *inferred* to be a property of the other. The quesion is: Does such a definition of identity work in a decentralized environment such as the Web of Linked Data?

3 The Identity Crisis of Linked Data

Just because a construct in a knowledge representation language is explicitly and formally defined does not necessarily mean that people will follow that definition when actually using that construct 'in the wild.' This can be for a wide variety of reasons. In particular, the language may not provide the facilities needed by people as they actually try to encode knowledge, so they may use a construct that appears to be *close enough* to what they need. A combination of not reading specifications—especially formal semantics, which even most software developers and engineers lack training in—and the labeling of constructs with "English-like" mnemonics, will naturally lead the use of a knowledge representation language by actual users to vary from what its designers intended. In decentralized systems such as the Semantic Web, this problem is amplified. Far from being a sign of abuse, it is a sign of success, as it means that the Semantic Web is actually being deployed outside academia and research labs.

[1] For example, using a *single* pre-defined criterion to define identity has been a success in terms of primary keys in databases. OWL also allows us to deploy such a property using the *owl:inverseFunctionalProperty* construct, although this is a rather simple approximation of a full-fledged theory of identity criterion.

At first glance, *sameAs* seems to be harmless. Its informal definition is that "the built-in OWL property *owl:sameAs* links an individual to an individual" and "Such an *owl:sameAs* statement indicates that two URI references actually refer to the same thing: the individuals have the same identity" [1]. OWL states that "It is unrealistic to assume everybody will use the same name to refer to individuals. That would require some grand design, which is contrary to the spirit of the web" [1]. The problems with *sameAs* start when we apply the principle of substitution to it, by inferring from a *sameAs* assertion that its subject and object share all the same properties.

Despite efforts such as OKKAM which attempt to get the Semantic Web to re-use URIs [4], with the distributed growth of Linked Data projects new URIs are often being minted for new data-sets independently and then *sameAs* links are added manually or automatically. Furthermore, the entire transitive closure of *all* individuals that are connected by *sameAs* share *all* the same properties, if the official (substitutive) definition is respected.

There is the possibility that *sameAs* could turn the Semantic Web from a web of interconnected data to the semantic equivalent of mushy peas. Of course identity is transitive and substitutive. If all the uses of *sameAs* are semantically correct, all these entailments would be exactly correct. The problem is not that *sameAs* itself is mushing up *Linked Data*, but that it's being used to mean other things than what the specification says it means.

While there have been heroic efforts to deal with these 'co-reference' bundles by the KnoFuss architecture [15] and the Consistent Reference Service [7], these have both been deployed only in certain domains. While there has been much related work in the database community on assessing information quality from uncertain sources of information [16], and some work in the Semantic Web community such as the work of WIQA [2] and Inference Web [13], this work has yet to be widely deployed for Linked Data. As imaginable, this has led to considerable discussion in the Linked Data community that such use of *sameAs* is dangerous and potentially 'wrong' as regards the formal semantics of OWL 1.0. However, since inference is rarely used with Linked Data, these problems are not always noticed. Does the possibility of incorrect inferences even matter if one's application does not use inference? With frameworks such as SiLK increasing the number of *sameAs* [17] statements, is the use of *sameAs* a potential time-bomb for Linked Data, or just a harmless convention?

4 Varieties of Identity and Similarity

What kinds of uses of *sameAs* inconsistent with its strict logical definition may be found in the wild? The kind of uses we find suggest that in some cases the context (which can be given on the Semantic Web as a named graph) of the use of name of is *referentially opaque* despite both names denoting a single thing. In other cases the two things are just similar. In neither case is it implied that either name can be freely substituted for the other (the Principle of Substitution

does not hold), nor can all the properties of either name be inferred to hold of the other.

4.1 Identical But Referentially Opaque

The first case is when things are **identical**, *that is the two names do identify to the same thing, but all the properties ascribed to one name are not necessarily appropriate for the other*, so their names can not be substituted. In this case, the context of use, like a named graph on the Semantic Web, is referentially opaque. While this may appear to violate the very definition of identity, there are two general cases where this may hold.

The first case is when indeed the two names do identify the same thing, but not all properties asserted using one of the names may be asserted using the other name. This is the case when the particular name used to refer to an object matters in some important way. A typical example of referential opacity arises when we have an explicit representation of an agent's knowledge or belief, and the agent doesn't know that the names co-refer. If the agent believes that the 'Morning Star' refers to Venus, but does not know that the 'Evening Star' also refers to Venus, then an equality substitution (such as using *sameAs*) between the 'Evening Star' and 'Morning Star' will give a false representation of their beliefs, even though this equation is factually true.

Another case is when two names may refer to the same thing and all properties do hold of both names, but it is socially inappropriate to re-use the name in a different context (a context can be given as a named graph in RDF). The central intuition here is there are 'forms of reference' appropriate to a context, especially in social contexts. To use an informal example, when at an event of the Royal Society, Tim Berners-Lee is *Professor Sir* Tim Berners-Lee of MIT and Southhampton, not *timbl* on IRC. This does not mean that in an IRC chat Tim Berners-Lee is somehow *not* a professor, but that within that context those properties do not matter. This property is exceedingly important for Linked Data, as contrary to popular doctrine, URIs are uused often as kinds of names and it is possible that the Web is full of referentially opaque contexts.

4.2 Identity as Claims

One could attempt to avoid the entire problem by simply treating all statements of identity as **claims**, *where the statement of identity is not necessarily true, but only stated by a particular agent.* As different agents may have different sets of claims they accept, different agents may accept different identity statements and so have different inferences. These issues also apply to the Semantic Web insofar as it uses any kind of inference as once an agent accepts an identity claim, the agent is bound to all its valid inferences. Informally, it is one thing for me to link to your URI, but its another thing for me to believe what you say about it as though you were talking about my URI. Put another way, one should be wary of accepting conclusions *over here* that could have been drawn *over there*, so to speak.

In particular, this issue comes into play when different agents describe the world at different levels of granularity. For example, different sources of Linked Data may make subtley different claims about some common-sense term like 'sodium.' This occurs in the case of the concept of sodium in DBPedia, which has a *sameAs* link to the concept of sodium in OpenCyc. The OpenCyc ontology says that an element is the set of all pieces of the pure element, so that sodium in Cyc has a member which is a lump of pure metallic sodium with exactly twenty-three neutrons. On the other hand, sodium as defined by DBPedia includes all isotopes, which have different number of neutrons than 'standard' sodium, and in this particular case are unstable. So, one should not state the number of neutrons in DBPedia's use of sodium, but one *can* with OpenCyc. At least in web settings with little inference or reliance on detailed structures, it is unlikely that most deployers of Linked Data actually check whether or not *all* the properties and their associated inferences are shared amongst linked data-sets.

4.3 Matching

As inspired by *skos:exactMatch*, which states "indicates a high degree of confidence that two concepts can be used interchangeably across a wide range of information retrieval applications." [14], one can imagine a kind of strong similiarity relationship called **matching** where *different things share enough properties enough to substitute for each other*, at least for some purposes. Unlike *skos:exactMatch* this property would apply to things themselves, not just concepts of things. Two descriptions of things can share all the same properties due to only a finite and incomplete number of these properties being described. For example, while a wine-glass is identical to itself, it would match another wine-glass from the same set in a Semantic Web description...at least for the purposes of laying a table. We should also be careful not to mix up names and things. The "Department of Paris" and "District of Paris" may share the same geographical extent, but by what act of civil engineering on a grand scale or legal act in court could such things actually be substituted for each other? Obviously they are not identical and only strongly similar, even if the knowledge representation of them lists all the same properties by virtue of being incomplete.

4.4 Similar

Another relationship is a kind of weaker notion of being **similar**, which is when *two different things share some but not all properties* in their given incomplete description. A wine glass and a coffee-cup are similar as regards holding liquids, but they hold entirely different kinds of liquid usually and are different shapes, so Leibnitz's Law would not hold obviously as they are different things. A real-world example from Linked Data would be the relationship between two biospecimens coming from the same cell line in an experiment [12]. We have observed scientists inclination in practice to connect them with *sameAs*, as the two biospecimens are part of the same cell line. However, this creates inferential problems including causing the specimen to be derived from itself, and important experimental

properties to be duplicated! Therefore, it makes more sense to have an identifier that only causes some (but perhaps not all) properties to be shared.

4.5 Related

The final relationship is **related**, when *two different things share no properties in common in a given description but are nonethless closely aligned in some fashion.* For example, the relationship between a wine-glass and wine. Such complex, structured, yet hard-to-specify relationships between things that are 'kind of close to identity' often arise, such as the relationship between a quantity and a measurement of a quantity and between sodium and a isotope. One example of this from Linked Data is the use of a drug in a clinical trial and the drug itself, which is currently connected via *sameAs* in a Linked Data drug study [10]. Although on some trivial level 'everything is related', there are degrees of relatedness. A drug may be related to many things (such as certain plants it derived from), that fact may have little relevance to, much less identity with, the clinical trial that tested its properties, as these properties could also be synthetically brought about. One is also tempted to engage with some sort of "fuzzy" or numerical weighted uncertainty measure between one and zero of identity, but the real hard questions of precisely where these real values come from and their relationship to actual probability theory muddy these conceptual waters very quickly. It seems that beneath these predicates there is likely to be a whole family of heterogeneous and semi-structured relationships that should be studied more carefully and empirically observed before any hasty judgments are made.

5 The Similarity Ontology

Although in Section 4 we demonstrate a need for a notion of identity that does not have any entailments and the possibility that various forms of similarity are being confused with the notion of identity, we did not explicitly explore the details. One possibility as originally proposed and discussed in [12] would be to propose a number of new relationships of identity based on permutations around each of the properties of transitivity, symmetry, and reflexivity. A new ontology called the *Similarity Ontology* (SO) has been defined that separates each of these out as a new kind of relationship.[2] While one could use these properties to make inferences about the relationship in certain domain-specific cases, one would not thereby necessarily be claiming that any two objects having this new kind of relationship would share properties.

The properties of the Similarity Ontology are shown in Table 1. Unlike identity, similarity properties are not necessarily transitive and symmetric. Note that non-symmetric is not equivalent to asymmetric, but simply not necessarily symmetric. The same applies to non-reflexivity and irreflexivity, and non-transitivity

[2] http://purl.org/twc/ontologies/similarity.owl

and intransitivity. Domain-specific properties can be created as sub-properties of one of the eight SO properties in order to maximize interoperability while maintaining distinctions among future concepts of similarity. We have also defined a mapping ontology that shows examples of mappings with existing similarity properties from RDFS, OWL, and SKOS[3] and show the sub-property relationship among the new and existing similarity properties in Fig. 1. These properties cover the wide range of relationships from "a is the same thing as b" to "b has more information about a" and allow the expression of precise concepts of similarity.

		Transitive	Non-transitive
Reflexive	Symmetric	*so:identical*	*so:similar*
	Non-Symmetric	*so:claimsIdentical*	*so:claimsSimilar*
Non-Reflexive	Symmetric	*so:matches*	*so:related*
	Non-Symmetric	*so:claimsMatches*	*so:claimsRelated*

Table 1. The proposed Identity Ontology. Eight new identity properties derived from the original meta-properties of *sameAs*: Reflexivity, Symmetry, and Transitivity. The prefix "sim" is used for the ontology.

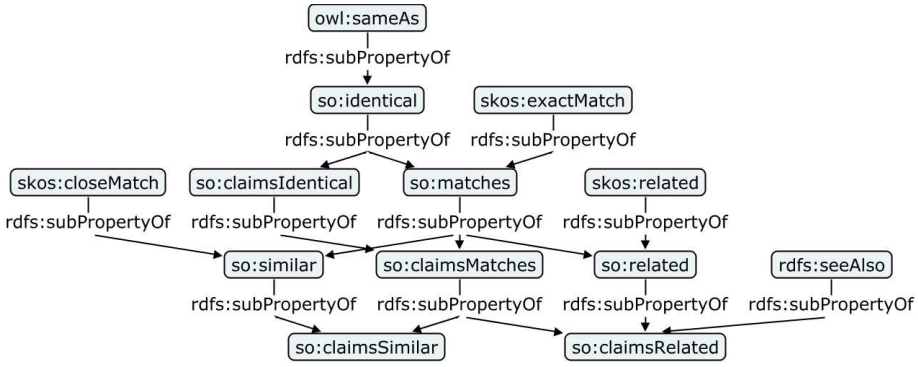

Fig. 1. Sub-property relationships between the properties of the Similarity Ontology and existing properties from OWL, RDFS, and SKOS.

so:identical Two URIs refer to the same thing and so share all the properties, but the reference is opaque. This is the most restrictive property of similarity in SO. It follows the same definition as *sameAs*, which "indicates that two URI references actually refer to the same thing: the individuals have the same identity", but it is referentially opaque and so does not follow Leibnitz's Law

[3] http://purl.org/twc/ontologies/similarity-mapping.owl

[1] As this is the most restrictive property, no other SO properties are sub-properties of it. *sameAs* is defined to be a sub-property so that existing valid assertions of identity are preserved.

so:claimsIdentical Since this property is transitive and reflexive, but not necessarily symmetric, it serves as a way for one agent to claim two URIs are identical, without the inverse needing to be true. As a super-property of *so:identical*, everything that is actually identical makes the claim of identity, with both sides of the claim being made due to the symmetry of *so:identical*. This property is transitive because if an entity a claims to be entity b and b claims to be entity c, then a cannot deny that it is claiming to be c as well.

so:matches Two URIs refer to possibly distinct things that share all the properties needed to substitute for each other in some graphs. This property is symmetric but not necessarily reflexive. *so:matches* is a super-property of *so:identical*.

so:claimsMatches This is the same as *so:matches*, but is not necessarily symmetric, so that things can be claimed to match without reciprocation.

so:similar Two URIS refer to possibly different things that share some properties but not enough to substitute for each other. *so:similar* is a super-property of *so:matches*. This is a super-property of *so:identical* since everything that is identical is also similar. It is also a super-property of *skos:closeMatch*[14].

so:claimsSimilar This is the same as *so:similar* but is not necessarily symmetric. Agents can therefore use this property to claim similarity without reciprocation. As a statement of similarity is in actuality two claims of similarity, so *so:claimsSimilar* is a super-property of *so:similar*. In symmetry with *so:similar*, claims of identity and matching imply a claim of similarity.

so:related Two URIS refer to possibly distinct things, and share no properties necessarily but are associated somehow. As it is only symmetric, there are no claims to any sort of similarity, matching, or identity. Because of this, *so:related* is a super-property of only *so:matches*, as *so:similar* and *so:identical* are reflexive, which would make *so:related* reflexive by proxy. This property is closely related to *skos:related* [14].

so:claimsRelated This is the loosest sense of identity in SO. It is a similar property to *rdfs:seeAlso*, which is "used to indicate a resource that might provide additional information about the subject resource." [5] We define *rdfs:seeAlso* to be a sub-property of *so:claimsRelated*. *so:related* and *so:claimsMatches* are both super-properties of *so:claimsRelated*.

5.1 Inference

There is a real opportunity here for doing inference. How is this done? It can be said that a particular property or set of properties are isomorphic across a particular kind of similarity. This kind of entailment can be performed through introduction of a property chain, introduced in OWL 2. What people obviously want to express is 'same cell line as,' or more generally, 'same relevant property as' (One could imagine a number of relevant properties and sub-properties). This is much more structured than a vague notion of matching and similarity,

and probably more useful. We could do this in OWL now by having a class of identity-restrictions, along these lines:
sameAsClass a IDRestriction.
samePropertyAs relevantProperty P.
A samePropertyAs B.
A P X. B P Y. $\rightarrow$
X sameAs Y.

6 Experiment

We have carried out an empirical study of *sameAs* "in the wild". Examples of *sameAs* were taken from the Linked Data Web in order to determine how robust the distinctions offered above are in practice. That is, do people actually use *sameAs* in the different ways that are outlined in the Similarity Ontology? Can people recognize these kinds of distinctions reliably? If at least some of the distinctions between similarity relationships that are currently conflated by *sameAs* can be made in a robust manner, then these distinctions may be candidates for standardization.

6.1 Data

For our experiment we retrieved all *sameAs* triples from the copy of the Linked Open Data Cloud hosted by OpenLink, which totalled 58,691,520 *sameAs* triples from 1,202 unique domain names. The top eight providers of triples show a heavy slant towards biology, being in order: *bio2rdf* (26 million), *uniprot* (6 million), DBPedia (4.3 million), Freebase (3.2 million), Max Planck Institute (.85 million), OpenCyc (.2 million), Geonames (.1 million), Zemanta (.05 million). As shown in Figure 2, when the domain of each URI in the subject and object is plotted by rank-frequency in log-log space, these triples display what appears to be power-law behavior. This is in line with earlier results [8] that show that Linked Data does not necessarily follow a power-law, but something relatively close that does exhibit a somewhat fore-shortened long-tail and nearly exponential behavior in the head. When we used the standard method of Clauset et al. to detect a power-law, the exponent was estimated to be 2.42, but the Monte-Carlo generation of synthetic distributions showed that the distribution failed significantly ($p = .08, p \leq .1$, no power-law found) to be a power-law. Nonetheless, it is seemingly exponential and almost certainly non-parametric.

In order to select a subset for an initial experiment, we first eliminated some classes of triples, and then took a weighted random sample. As the data was to be rated by non-specialists, all biomedical data with *bio2rdf* and *uniprot* links was excluded from the random sampling. Furthermore, the two linked data-sets that just copied data (DBPedia) blindly, namely *zemanta* and *Freebase*, were also excluded.

We then drew approximately 500 sample *sameAs* statements at random from the remaining 2.3 million triples. In order to prevent the major data-providers

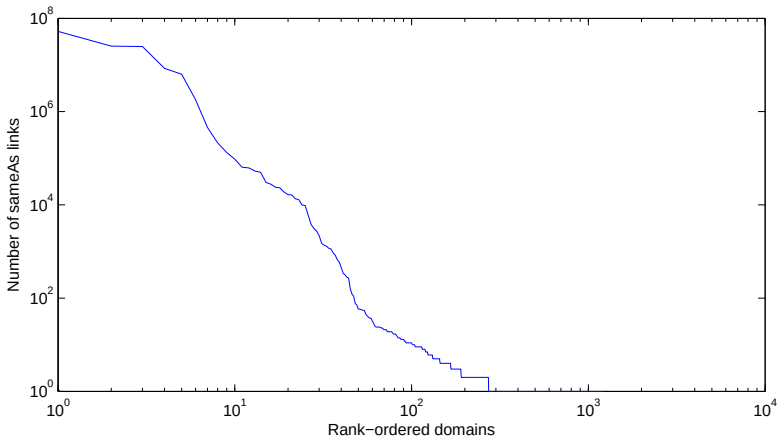

Fig. 2. Frequency of domains in *sameAs* statements in rank-order, logarithmic (base 10) scale.

from unfairly dominating the sample, the samples were chosen so that the frequency of URIs in the resulting triples from major providers (those in the exponential head of the distribution) was scaled down by the logarithm of their raw frequency. This down-weighting is intended to result in a balanced and diverse sample of *sameAs* statements. Finally, we attempted to retrieve RDF triples whose subjects were the subject and object URIs of those statements. The 250 cases where this retrieval was successful provided the material for our initial evaluation experiment.

6.2 Experimental Design

We used the *Amazon Mechanical Turk*[4] as a platform for a pilot experiment. Tasks that require some amount of human judgement (such as the judgement about identity) are broken into what are termed Human Intelligence Tasks (HITs) for presentation via the Web to three of the authors. Each HIT covered 10 *sameAs* pairs, as shown in Figure 3, with a standard sample of properties and values from each retrieved RDF triple displayed in two side-by-side tables. We hope to later repeat this experiment on a larger scale using crowd-sourcing via this platform.

The following instructions were given for the forced choice response: **The same:** clearly intended to identify the same thing, without necessarily using the same properties e.g. two different descriptions of a live performance by Queen of *Bohemian Rhapsody*. **Matches:** identifies two copies or versions of the same thing, with the same fundamental properties and differing only with regards to incidental properties, e.g. descriptions of two live performances by Queen of

[4] *https://www.mturk.com/*

Bohemian Rhapsody, but at different locations. **Similar:** Identifies two fundamentally distinct things, but with some properties in common e.g. descriptions of two live performances of *Bohemian Rhapsody*, by two different bands. **Related:** not intended to identify the same thing, but related. E.g. descriptions of the band Queen and of a live performance by Queen of *Bohemian Rhapsody*. **Unrelated:** None of the above. Also, a 'Can't tell' response was available.

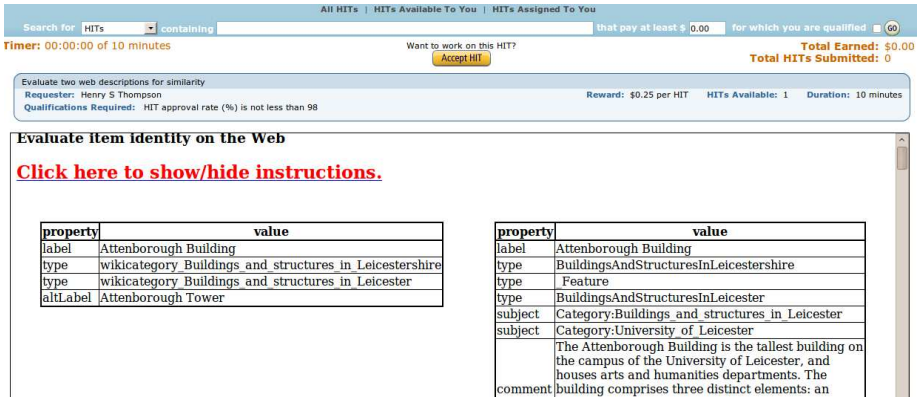

Fig. 3. Mechanical Turk Interface for identity rating.

As a step towards creating a gold standard, three of the authors assessed all 250 samples. We plotted the results for each judge per category in Figure 4, revealing what appears to be substantial disagreement with respect to some categories. Merging the results of each judge, a table is given in Table 2 that gives raw agreement and disagreement frequencies.

First of all, the vast majority of *sameAs* statements were indeed judged to be correctly identical, and only a relatively small amount were judged to be incorrect. Interestingly enough, a relatively large amount were unknown. Only a small amount were judged as similar, while the amount judged to be matches and related were modest. To return to Figure 4, it is very clear that the judges have different styles of judgement, with one judge preferring *sameAs* where another judge would be much more strict by usually answering that they can't tell. The remaining judge is in between these two extremes. The amount of disagreement shows that the categories are fairly unstable. However, there is clearly *something* in between not knowing if two URIs are identical and knowing that they are.

Since each question could be considered a binary response over nominal data, we employed the κ statistic to determine agreement between the judges. The κ statistic takes into account agreement between annotators that is greater than chance, and is only valid over nominal data (although our data could be considered ordinal, it is strictly speaking nominal, as each choice is a different relation-

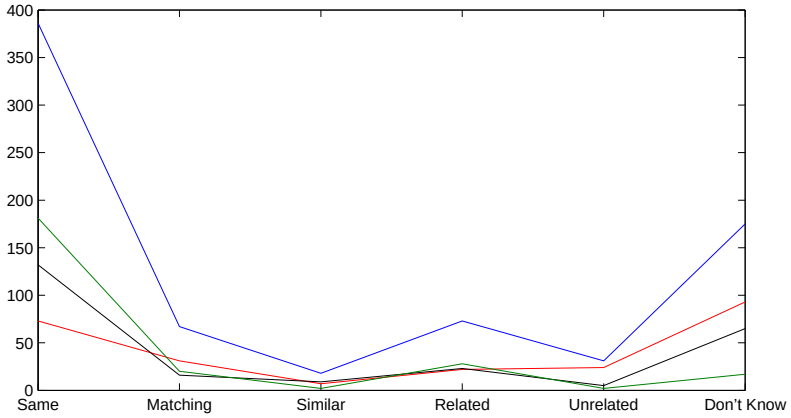

Fig. 4. Number of category assignments per judge. Total across all judges blue, each individual judge is red (1), black (2), and green (3). Y-axis is their frequency in the data-set.

ship rather than a single identity gradient).[5] The κ for the six-way forced choice is 0.158, which is non-accidental but considered 'poor' agreement. Notice that while there was substantial disagreement, there were elements (particularly of identical) where nearly half the data-set was labelled in agreement, likewise for the 'related' category and a substantial portion of 'don't know'. However, the rest of the categories appear to be terminally prone to error. By optimizing and recombining categories, we were able to reach a κ of .319, which indicates 'fair' agreement. This was accomplished by merging the 'similar', 'matching', and 'related' categories, and then merging the 'can't tell' with 'not related' categories, and leaving the 'same as' category to itself. The results, as given per judge in Figure 5, are much more clear. However, there is still substantial disagreement. The main disagreement seems to consist of, rather surprisingly, an inability to agree on 'same as' versus 'can't tell'.

Categories-Rater	Rater 1	Rater 2	Rater 3
Identical	73	132	181
Matching	31	16	20
Similar	7	9	2
Related	22	23	28
Not Related	24	5	2
Can't Tell	93	65	17

Table 2. Raw numbers of Similarity Categories before optimization.

[5] The derivation of the κ statistic is described in mathematical detail elsewhere [6].

The differing habits of the raters in this regard are actually more unstable than their ability to link something using a 'sort of similar or related' category, as shown by inspection of Table 3. It is not in the categories themselves that the problem surfaces, but in the lack of appropriate knowledge for use in determining whether two things are in some context-free manner actually identical. This brings into some doubt the concept of whether or not two things can be declared identical in a context-free manner, and also highlights the importance of background knowledge in determining accurate *sameAs* statements. In this regard, it should not be surprising that there was such high disagreement on manual judging of identity and similarity in Linked Data. However, there are a number of positive results that we can make a guess at by taking the mean of the collapsed categories per rater (and their standard deviation):

- The most postive result is that approximately 51% (± 21%) percent of the usage of *sameAs* seems correct.
- While the distinctions made in the Similarity Ontology likely require special training beyond that of even RDF experts, a relatively coarse-grained referentially opaque 'kind-of-similar-and-related-to' relationship can be reliably used instead of *sameAs* for intermediate cases (around 21% (± 3%) of our data);
- Approximately 27% (± 19%) of the *sameAs* cannot be reliably judged based only on the RDF retrieved.

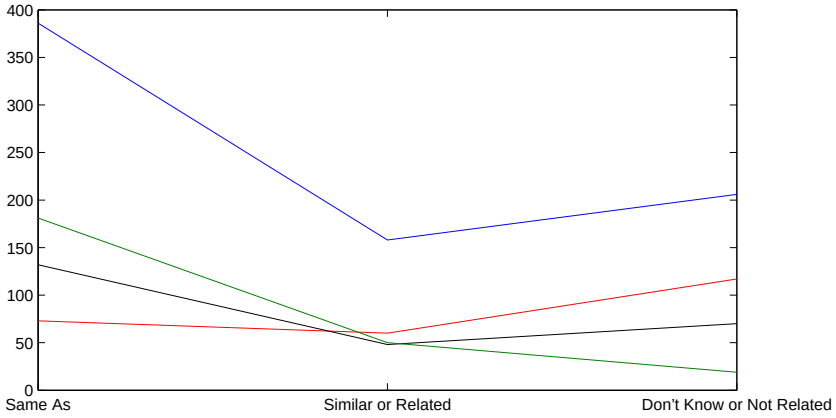

Fig. 5. Frequency of categories in trained expert judges after optimization. Total across all judges blue, each individual judge is red (1), black (2), and green (3). X-axis is categories, Y-axis is their frequency in the data-set.

Categories-Rater	Rater 1	Rater 2	Rater 3
Identical	73	132	181
Similar, Matching, Related	60	48	50
Can't Tell or Not Related	117	70	19

Table 3. Raw numbers of Similarity Categories after optimization.

7 Conclusion

The issue of how to express relationships of identity and similarity on Linked Data is more complex than just applying *sameAs*. We believe the extent of disagreement and inaccurate usage as observed in practice at least calls for additional documentation providing clearer guidance on when to use *sameAs*. Further studies on much larger scales using crowd-sourcing need to be employed to see if the 'default' behaviors of the judges in our experiment generalizes. A further extension of our experiment will test whether the closures of *sameAs* produce surprising and incorrect inferences. This can be done by merging inferred triples with the *sameAs* statements used in the current experiment.

The proposed Similarity Ontology solution has a number of distinctions that may be difficult to deploy consistently in open-ended domains. In fact, like many ontologies, the initial distinctions proposed capture an important intuition, namely that there is a nuanced heterogeneous structure of similarity instead of a strict notion of identity in the use of *sameAs* on the Web, one that will likely result in an asymmetric flow of inference. However, the Similarity Ontology explores too large of a design space to be reliably deployed. A simple similarity property would be quite useful to add to RDF, such as sub-property of *rdfs:seeAlso*. Further study of approaches beyond *sameAs* would be useful if not provocative for the Linked Data community. Solving the issue of identity in Linked Data may require a certain refactoring of some core constructs of RDF, including relating identity to a fully-worked out semantics for named graphs. Furthermore, individuals could be thought of as being composed of differing aspects at different levels of granularity rather than the notion of individuals traditionally used in semantics. In future work, we will also continue investigations into the notion of aspects and named graphs and continue to be inspired by the use cases presenting themselves from the current abundance of misuse of *sameAs* in Linked Data space. The (ab)use of *sameAs* in Linked Data is not a threat, it's an opportunity.

8 Acknowledgements

We would like to thank reviewers of earlier versions of this work in OWLED 2010, LDOW 2010, and RDF Next Steps for their helpful feedback. Also, special thanks to Kingsley Idehen for helping provide the data-set.

References

1. S. Bechhofer, F. van Harmelen, J. Hendler, I. Horrocks, D. L. McGuinness, P. F. Patel-Schneider, and L. A. Stein. OWL Web Ontology Language Reference, 2004.
2. C. Bizer and R. Cyganiak. Quality-driven information filtering using the wiqa policy framework. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(1):1 – 10, 2009.
3. C. Bizer, R. Cygniak, and T. Heath. How to publish Linked Data on the Web, 2007. http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/ (Last accessed on May 28th 2008).
4. P. Bouquet, H. Stoermer, and D. Giacomuzzi. OKKAM: Enabling a Web of Entities. In *I3: Identity, Identifiers, Identification. Proceedings of the WWW2007 Workshop on Entity-Centric Approaches to Information and Knowledge Management on the Web, Banff, Canada, May 8, 2007.*, CEUR Workshop Proceedings, ISSN 1613-0073, May 2007. online http://CEUR-WS.org/Vol-249/submission_150.pdf.
5. D. Brickley and R. V. Guha. RDF Vocabulary Description Language 1.0: RDF Schema, 2004.
6. J. Carletta. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22:249–254, 1996.
7. H. Glaser, I. Millard, and A. Jaffri. RKBExplorer.com: A knowledge driven infrastructure for Linked Data providers. In *Proceedings of European Semantic Web Conference (ESWC)*, pages 797–801, Tenerife, Spain, 2008.
8. H. Halpin. A query-driven characterization of linked data. In *Proceedings of the Linked Data Workshop at the World Wide Web Conference*. Madrid, Spain, 2009.
9. H. Halpin and P. Hayes. When owl:sameas isn't the same. In *Proceedings of the WWW2010 Workshop on Linked Data on the Web, Raleigh, USA, April 25th, 20010.*, April 2010. online http://events.linkeddata.org/ldow2010/papers/ldow2010_paper09.pdf.
10. A. Jentzsch, O. Hassanzadeh, C. Bizer, B. Andersson, and S. Stephens. Enabling tailored therapeutics with linked data. In *Proceedings of the WWW2009 Workshop on Linked Data on the Web, Madrid, Spain, April 20th, 20010.*, April 2009. online http://events.linkeddata.org/ldow2009/papers/ldow2009_paper9.pdf.
11. G. Leibniz and L. Loemker. *Philosophical papers and letters*. Springer, 1976.
12. J. McCusker and D. McGuinness. Towards identity in linked data. In *Proceedings of OWL: Experience and Directions, San Francisco, USA, June 21-22nd, 20010.*, June 2010. online http://www.webont.org/owled/2010/papers/owled2010_submission_12.pdf.
13. D. L. Mcguinness and P. P. Silva. Explaining answers from the semantic web: The inference web approach. *Journal of Web Semantics*, 1:397–413, 2004.
14. A. Miles and S. Bechhofer. SKOS Simple Knowledge Organization System Reference, 2009.
15. A. Nikolov, V. Uren, and E. Motta. Knofuss: a comprehensive architecture for knowledge fusion. In *K-CAP '07: Proceedings of the 4th international conference on Knowledge capture*, pages 185–186, New York, NY, USA, 2007. ACM.
16. L. L. Pipino, Y. W. Lee, and R. Y. Wang. Data quality assessment. *Communications of the ACM*, 45(4):211–218, 2002.
17. J. Volz, C. Bizer, M. Gaedke, and G. Kobilarov. Discovering and maintaining links on the web of data. In A. Bernstein, D. R. Karger, T. Heath, L. Feigenbaum, D. Maynard, E. Motta, and K. Thirunarayan, editors, *International Semantic Web Conference*, volume 5823, pages 650–665. Springer, 2009.