# Towards Identity in Linked Data

James P. McCusker and Deborah L. McGuinness

Tetherless World Constellation
Department of Computer Science
Rensselaer Polytechnic Institute
110 8th Street Troy, NY 12180, USA
{mccusj,dlm}@cs.rpi.edu
http://tw.rpi.edu

**Abstract.** Many Linked Data applications have come to rely on owl:sameAs for linking datasets. However, the current semantics for owl:sameAs assert that identity entails isomorphism, or that if a=b, then all statements of a and b are shared by both. This becomes problematic when dealing with provenance, context, and imperfect representations, all of which are endemic issues in Linked Data. Merging provenance can be problematic or even catastrophic in biomedical applications that demand access to provenance information. We use examples in biospecimen management, experimental metadata representations, and personal identity in Friend-of-a-Friend (FOAF) to demonstrate some of the problems that can arise with the use of owl:sameAs. We also show that the existence of an isomorphic owl:sameAs can be inconsistent with current expectations in a number of our use cases. We present a solution that allows the extraction of isomorphic statements without requiring their direct assertion. We also introduce a set of identity properties that can be extended for domain-specific purposes while maintaining clarity of definition within each property.

**Key words:** owl:sameAs, identity, linked data, inferencing

## 1 Introduction

The adoption of Linked Data by many communities has resulted in a wealth of useful data that can be combined in many ways. One current practice in Linked Data, and therefore in many resources on the Semantic Web, is to link datasets using *owl:sameAs* to define that two entities between those datasets are the same. On its face, this is a good idea because a higher level best practice is to re-use properties when possible and appropriate and *owl:sameAs* is a property that seems easy to understand and is part of OWL [1], which is one of the foundations of the semantic web and Linked Data.

The property *owl:sameAs* is a very strict notion of identity. Its definition stems from that of mathematical identity, most specifically, from the Indiscernibility of Identicals [2], or $a = b \wedge p(a, x) \Rightarrow p(b, x)$. This law is true for all true statements about $a$ and is the basis for *isomorphism* in *owl:sameAs*. Halpin and

Hayes [3] identify four different uses of *owl:sameAs* in Linked Data: (1) Same Thing As But Different Context, (2) Same Thing As But Referentially Opaque, (3) Represents (4) Very Similar To. They question if the use of *owl:sameAs* in Linked Data is truly a harmless convention. Some have argued that since Linked Data applications rarely use inference, one should not worry about computational problems that would result from *owl:sameAs* inference. We argue that at least some types of Linked Data applications (such as those shown in our examples) do benefit from and actually require inference.

We argue that one major issue with the use of *owl:sameAs* in Linked Data and the Semantic Web is related to the inference of isomorphism. We present some biomedical examples that help to describe some problems with *owl:sameAs* and isomorphic inference. Finally, we discuss a set of possible identity properties and show how the current uses outlined by Halpin and Hayes and other examples that we describe fit into the new framework.

## 2 Problems with Identity in Linked Data

We present some examples that help to describe problems with *owl:sameAs* and isomorphic inferencing. For instance, Jaffri *et al.* discusses a number of erroneous mappings of *owl:sameAs* in DBpedia and DBLP [4]. Ding *et al.* [5] discusses a number of issues around combining FOAF profiles using *owl:sameAs*. Finally, Jain *et al.* [6] argues that more expressive semantics should be used to improve the richness of Linked Open Data (LoD) Cloud.

There are also issues with creating Linked Data while maintaining provenance information. An example that we have found is in biomedical data. In Fig. 1, we show a common derivation graph of a tumor $T$ that was removed from a patient $A$. A cell line was derived from the tumor which resulted in two specimens, $LB$ and $LA$, where $LA$ was the original cell colony and $LB$ was derived from it by taking a part of $LA$ and growing it on its own. There are datasets $(D, E)$ that use $(LB, LA)$ respectively, and a scientist would like to integrate those datasets. Through discussions with practicing scientists, we became aware of situations where *owl:sameAs* is used to accomplish this. For example, the use of *owl:sameAs* in Fig. 2 allows a scientist to integrate the datasets shown. However, isomorphism is applied to all other properties of the specimens, resulting in ambiguous or contradictory information. For instance, both specimens are inferred to have been created on both 8/31 and 9/20, and have a quantity of both 5 and 10 grams. Also, specimen $LA$ now seems to be derived from itself. It has been noted that these properties could be set as annotations, which would mean that they are not subject to isomorphism. However, most of the properties specified here are first-class data as specified by biospecimen management systems and would not be considered annotations by the originating system. There are other issues with this sort of inference. It is now very difficult to address the following potential problems that can arise:

– The data doesn't look right. What were the methods and protocols, and how consistent were they, going back to surgical resection?

– Did the "same cell line" actually come from the same tumor, or just from the same patient? Or even different patients?
– What originally seemed to be a primary breast cancer or lung cancer is now a metastasized melanoma. How do we sort this out?
– Is a histology slide made from a tumor the same as the tumor? What about the tissue microarray, the cell culture, or the isolated molecular material?

None of these issues are problems that always come up, but when they do, it is critical that the provenance of biospecimens remains distinguishable.

Some have argued that these problems can be surmounted by putting the offending statements in a named graph, or by attaching provenance statements as OWL 2 annotations on individuals. These solutions may work for limited contexts, but there are issues with both. Relying on named graphs is problematic because the statements may be embedded in other datasets which are already enclosed in named graphs. The user would have to extract identity statements from the original named graph and move them to another. Using annotations would work well if the problem were limited to provenance, and provenance were easily segregated into annotation properties. However, one data model's provenance is another's data. Biospecimen management systems concern themselves almost exclusively with what would be considered provenance to the experimentalist and analyst. It must be possible to realize *ex post facto* that large sections of a data model are considered to be provenance by other users. As we will show, this problem is not confined to provenance, but applies more generally to statements of knowledge. Neither of these solutions address these issues.

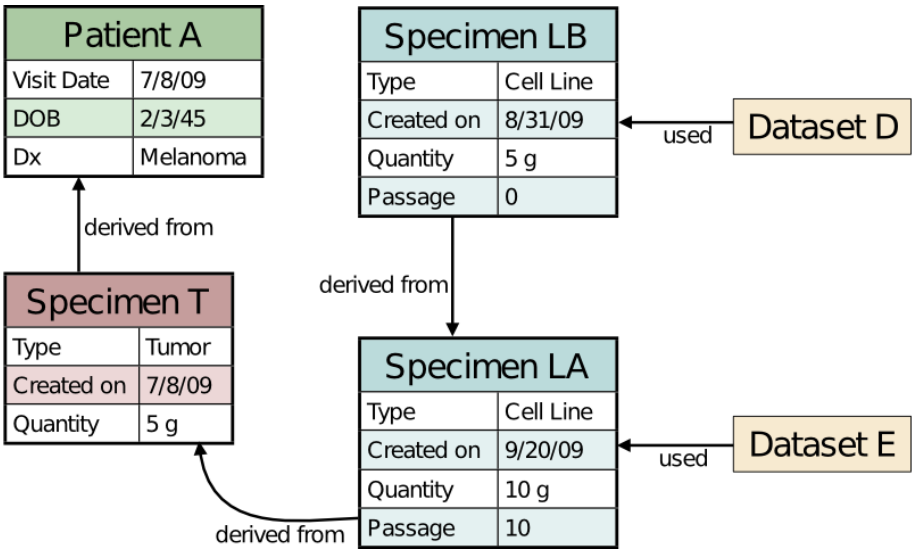


**Fig. 1.** A scientist has datasets $D$ and $E$, but $D$ and $E$ refer to different instances of the same cell line.

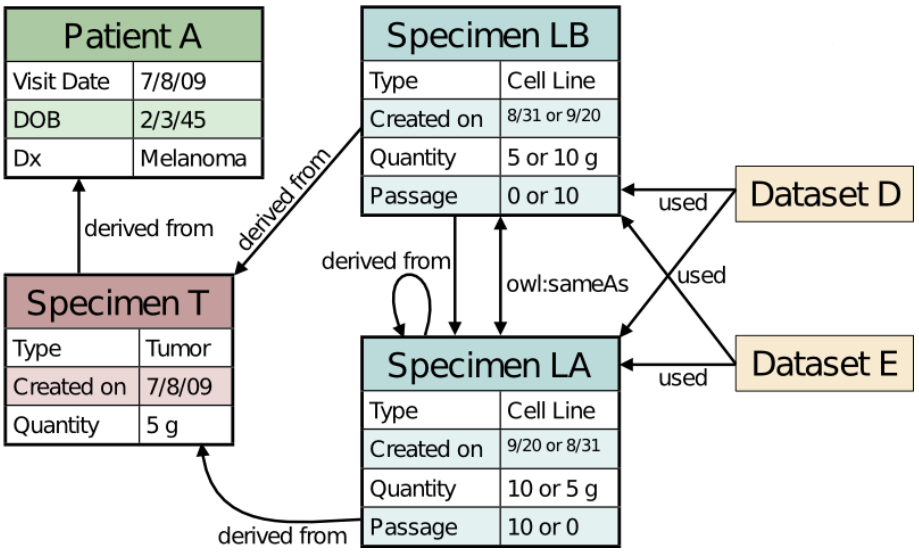


**Fig. 2.** As in Fig. 1, but with the assertion *owl:sameAs(LA,LB)*, then $D$ and $E$ can be integrated because they can refer to the same specimens. However, doing so means that there are now multiple values for some important properties and $LA$ appears to have been derived from itself.

## 3 A Hypothesis About the use of *owl:sameAs* in the Semantic Web

There are some additional problems with using the indiscernibility of identicals in knowledge representation discussed by Saul [7]. While the question of exactly what knowledge is is an ongoing philosophical problem [8, 9], it is generally agreed upon that for an entity to know a statement, it must first believe that statement to be true. This of course doesn't imply that the statement is true. False beliefs are not by any means rare. Because of this, a number of issues that apply to belief also apply to epistemology and knowledge representation. Since the semantic web is concerned with knowledge and not truth values [10–12], these issues also apply to the semantic web. Below, we discuss some problems described by Kripke [13], Saul [7], and Pitt [14], but convert beliefs into assertions, which are more applicable to a semantic web context. One classic problem concerns secret identities:

1. Lois Lane claims Superman can fly.
2. Lois Lane claims Clark Kent cannot fly.
3. Superman and Clark Kent are the same person.

We can then infer that Lois Lane claims that Superman cannot fly, and that Clark Kent can. These statements in combination with the assertions generate contradictions. Similar problems arise even without belief statements:

4. Clark Kent is Superman's secret identity.
5. Therefore, Superman is Superman's secret identity.

This is nonsensical, as a secret identity must be different from a public identity. This is also true of changes in identity over time:

6. I never made it to Constantinople, but I visited Istanbul last week.
7. Istanbul was Constantinople.
8. I never made it to Istanbul, but I visited Istanbul last week.

These examples all violate Leibniz's law regarding the identity of indiscernibles [2] in various ways, showing that even things that seem identical may not be so. Another possibility is that a person's knowledge of $X$ is not an inherent predicate of $X$. This is important to the semantic web since, as we discussed before, all statements made in the semantic web are statements of knowledge, not statements of truth. Thus, it can be problematic to use *owl:sameAs* (with its isomorphic character) in all knowledge representation scenarios.

## 4 A New Model for Identity in the Semantic Web

The examples above indicate a need for an identity (or similar) notion that is not inherently isomorphic. We also have a need to be able to find all entities that are inferred (or stated) to be identical to a designated entity. The modeling we need to do that is driven by these and other use cases, lead us to desire more options related to the concept of identity. As mentioned above, Halpin and Hayes [3] have identified four usages of *owl:sameAs* in Linked Data and we have identified several more. We review owl:sameAs and decompose its properties into Transitivity, Symmetry, and Reflexivity. We then discuss different permutations of those properties and show how current usages of *owl:sameAs* fits into the new framework. Finally, we will show how isomorphism can be selectively enabled for particular properties using property annotations.

### 4.1 The Identity Ontology

The property *owl:sameAs* is, in addition to being isomorphic, Transitive, Symmetric, and Reflexive. Each permutation of those meta-properties can be viewed as a new kind of identity. We have defined a new ontology called the Identity Ontology (IO).[1] The properties of the IO are shown in Table 1. Domain-specific properties can be created as sub-properties of one of the eight IO properties in order to maximize interoperability while maintaining distinctions among future concepts of identity. We have also defined a mapping ontology that shows examples of mappings with existing identity properties from RDFS, OWL, and SKOS[2] and show the subproperty relationship among the new and existing identity properties in Fig. 3.

[1] http://purl.org/twc/ontologies/identity.owl
[2] http://purl.org/twc/ontologies/identity-mapping.owl

| | | Transitive | Intransitive |
|---|---|---|---|
| Reflexive | Symmetric | *id:identical* | *id:similar* |
| | Non-Symmetric | *id:claimsIdentical* | *id:claimsSimilar* |
| Non-Reflexive | Symmetric | *id:exactlyMatches* | *id:related* |
| | Non-Symmetric | *id:matches* | *id:claimsRelated* |

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

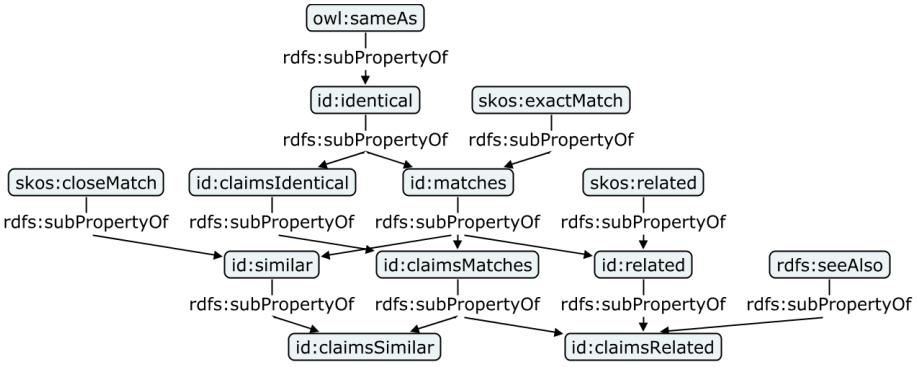


**Fig. 3.** Subproperty relationships between the properties of the identity ontology and existing identity properties from OWL, RDFS, and SKOS.

**id:identical** This is the most restrictive property of identity in IO. It follows the same definition as *owl:sameAs*, which "indicates" that two URI references actually refer to the same thing: the individuals have the same 'identity'." [1] As this is the most restrictive property, no IO identity properties are subproperties of it. *owl:sameAs* is defined to be a subproperty so that existing valid assertions of identity are preserved.

The usages "Same Thing As But Different Context" and "Same Thing As But Referentially Opaque" from Halpin and Hayes [3] fit neatly into *id:identical*. "Same Thing As But Referentially Opaque" is effectively supported directly by use of *id:identical*, and "Same Thing But Different Context" can be served by implementation of a subproperty to aid in distinction. The examples using Superman and Clark Kent can be considered to be in either class, and would be served equally well. The example using Istanbul[3] should be considered to be "Same Thing But Different Context", as the contextual distinction is the existence of the Ottoman Empire. The FOAF examples also would benefit from use of *id:identical*.

**id:claimsIdentical** Since this property is Transitive and Reflexive, but not Symmetric, it serves as a way for one entity to claim the identity of another,

[3] Not Constantinople.

without the inverse needing to be true. As a super property of *id:identical*, everything that is actually identical makes the claim of identity, with both sides of the claim being made due to the symmetry of *id:identical*. This property is transitive because if 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.

The usage "Represents" can be supported using *id:claimsIdentical* using a subproperty "representedBy". Since *id:claimsIdentical* suggests that $a$ can be replaced by $b$ if $claimsIdentical(a, b)$, then it can be said that $b$ represents $a$, or to fit into our usage more clearly, $representedBy(a, b)$.

**id:matches** This property is reflexive and symmetric. It is inspired by *skos:exactMatch*, which "indicates a high degree of confidence that two concepts can be used interchangeably across a wide range of information retrieval applications." [15] *id:matches* is a super property of *id:identical*, because for all things that are identical, they also match.

For extremely strong assertions of "Very Similar To" from Halpin and Hayes [3], *id:matches* can be used to assert identity because *id:matches* is intransitive. Many current identities in Linked Data would be well supported using this property.

**id:claimsMatches** This is the same as *id:matches*, but is not symmetric, so that entities can claim that they match things without reciprocation.

Weaker assertions of "Represents" can be supported using this method. It is also useful for representing the relationship between a particular biospecimen and the cell line that that represents it.

**id:similar** This is a statement of similarity without a guarantee of a complete match. Similarity is both Symmetric and Reflexive. Since things that match each other are also similar to each other, *id:similar* is a super property of *id:matches*. This is a super property of *id:identical* since everything that is identical is also similar. It is also a super property of *skos:closeMatch* [15].

This property can be best used to describe the fact that two biospecimens are part of the same cell line. A notional subproperty, such as sameCellLineAs, would allow for a domain-specific distinction of similarity that is understandable to domain experts while still providing usefulness to more general-purpose systems. Depending on the strength of "very" in "Very Similar To", it can also support the concept of identity for that usage from Halpin and Hayes [3].

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

This property is best seen in cases of asymmetric substitutions. For instance, decongestant can be substituted by an antihistamine (and can be said to be similar to it), but when someone has allergies, a decongestant will not relieve the symptoms. Another example is that, in a pinch, one can use conditioner in lieu of shaving cream, but the reverse does not hold.

**id:related** This asserts an associative link between two entities. As it is only symmetric, there are no claims to any sort of similarity, matching, or identity. Because of this, *id:related* is a super property of only *id:matches*, as *id:similar* and *id:identical* are reflexive, which would make *id:related* reflexive by proxy. This property is closely related to and is a super property of *skos:related* [15].

The idea of related entities is currently used in SKOS and in OBO (Open Biomedical Ontologies).[4]

**id:claimsRelated** This is the loosest sense of identity in IO. It is a similar property to *rdfs:seeAlso*, which is "used to indicate a resource that might provide additional information about the subject resource." [16] We define *rdfs:seeAlso* to be a sub property of *id:claimsRelated*. *id:related* and *id:claimsMatches* are both super properties of *id:claimsRelated*.

An example subproperty of *id:claimsRelated* is a depiction. Since a photograph or a illustration of a person or thing is not the thing itself, but a representation of the thing, this is a kind of identity that is not symmetric (the photograph is not depicted by the person), not transitive (a depiction of the depiction may not depict the original subject), and not reflexive (does a person depict themselves?).

These properties cover the wide range of identity relationships from "a is the same thing as b" to "b has more information about a" and allow the expression of precise concepts of identity while also leaving room for domain-specific concepts as well.

### 4.2 Reconstructing Isomorphism

For any reflexive statement of identity, it is possible to recover isomorphic statements using the following SPARQL query snippet:

```
select ?s, ?p, ?o
where {
    ?s id:identical ?x.
    ?x ?p ?o.
}
```

A major benefit of this formulation is that any property can be used in place of *id:identical* and can be used for domain-specific concepts of identity. Additionally, property chains in OWL 2 [17] allow the definition of isomorphism for

[4] http://obofoundries.org

specific properties where that behavior is warranted. The specific pattern would be:

$$SubObjectPropertyOf(ObjectPropertyChain(identical,\ p),\ p)$$

It is therefore possible to construct properties that are isomorphic across specific concepts of identity and allows users to query for values for any other property that would have been isomorphic if the identity had been asserted using *owl:sameAs*.

## 5 Discussion

We and others [3] have recognized the growing usage of OWL constructs such as *owl:sameAs*. However, we also have observed unanticipated usages of *owl:sameAs* where the existing semantics do not match the epistemological modeling needs. This has led us to develop the Identity Ontology. We believe that IO provides additional representational options for the notions of identity shown in our examples. We intend to continue our line of research into identifying, describing, and using these different representational options. It is interesting to note that the use cases are satisfied through application of existing OWL patterns of property types and property chaining.

The Identity Ontology is a starting point for developing a more nuanced approach to identity in the semantic web. IO addresses numerous challenges in our biomedical examples, and we have begun to use IO to represent concepts of identity in biomedical datasets. Specifically, we integrated two experiments from Array Express: E-TABM-65 and E-MEXP-1029. Both of these experiments use the NCI-60, a panel of cell lines used for cancer research. We converted the two experiments to RDF using MAGETAB2RDF[5] and aligned the biological sources using *biomedidentity:sameAsBioSource*.[6] In that ontology, we also make *mage:has_derivative*,[7] and *mged:has_biomaterial_characteristics*[8] isomorphic across our identity property using property chains. We plan to continue to investigate the properties of identity in relation to constructs included in IO as well as *owl:sameAs*.

## 6 Conclusions

We have elaborated on problems with isomorphism in the current use of *owl:sameAs* in Linked Data on the Semantic Web. We have provided more options for representing identity using our Identity Ontology and have initial work supporting its usage in a number of use cases. We also show how isomorphic statements can be queried, and how particular properties can be made to be isomorphic using property chains in OWL 2. We have also successfully used the Identity

[5] http://magetab2rdf.googlecode.com
[6] http://espresso.med.yale.edu/ jpm78/tw/identity/biomedidentity.owl
[7] http://magetab2rdf.googlecode.com/svn/trunk/ontologies/mage-om.owl
[8] http://mged.sourceforge.net/ontologies/MGEDOntology.owl

Ontology to enable more granular control over inference. We have found that this additional control over inferred information is a better match to our biomedical application needs than what we previously had access to using *owl:sameAs* alone.

## References


1. Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D.L., Patel-Schneider, P.F., Stein, L.A.: OWL Web Ontology Language Reference (2004)
2. Leibniz, G., Loemker, L.: Philosophical papers and letters. Springer (1976)
3. Halpin, H., Hayes, P.J.: When owl:sameAs isn't the Same: An Analysis of Identity Links on the Semantic Web. In: Workshop on Linked Data on the Web (LDOW). (2008)
4. Jaffri, A., Glaser, H., Millard, I.: URI Disambiguation in the Context of Linked Data. In: Workshop on Linked Data on the Web (LDOW). (2008)
5. Ding, L., Shinavier, J., Finin, T., McGuinness, D.L.: An Empirical Study of owl:sameAs Use in Linked Data. In: Web Science 2010. (2010)
6. Jain, P., Hitzler, P., Yeh, P., Verma, K., Sheth, A.: Linked Data Is Merely More Data. Dan Brickley, Vinay K. Chaudhri, Harry Halpin, and Deborah McGuinness: Linked Data Meets Artificial Intelligence. Technical Report SS-10-07 (2010) 82–86
7. Saul, J.: Substitution and simple sentences. Analysis **57**(2) (1997) 102
8. Armstrong, D.: Belief, truth and knowledge. Cambridge University Press London (1973)
9. Gettier, E.: Is justified true belief knowledge? Analysis **23**(6) (1963) 121
10. Berners-Lee, T., Hendler, J.: Scientific publishing on the semantic web. Nature **410** (2001) 1023–1024
11. Heflin, J.: Towards the semantic web: knowledge representation in a dynamic, distributed environment. PhD thesis, University of Maryland, College Park (2001)
12. Davies, J., van Harmelen, F., Fensel, D.: Towards the semantic web: ontology-driven knowledge management. John Wiley & Sons, Inc. New York, NY, USA (2002)
13. Kripke, S.: A puzzle about belief. Meaning and Use (1979) 239–283
14. Pitt, D.: Alter Egos and Their Names. The Journal of Philosophy **98**(10) (2001) 531–552
15. Miles, A., Bechhofer, S.: SKOS Simple Knowledge Organization System Reference (2009)
16. Brickley, D., Guha, R.V.: RDF Vocabulary Description Language 1.0: RDF Schema (2004)
17. Motik, B., Patel-Schneider, P.F., Cuenca Grau, B.: OWL 2 Web Ontology Language: Direct Semantics (2009)