owl:sameAs and Linked Data: An Empirical Study

Li Ding
Tetherless World Constellation
Rensselaer Polytechnic Institute
110 8th St., Troy NY 12180
dingl@cs.rpi.edu

Joshua Shinavier
Tetherless World Constellation
Rensselaer Polytechnic Institute
110 8th St., Troy NY 12180
shinaj@rpi.edu

Tim Finin
Computer Science and Electrical Engineering
University of Maryland, Baltimore County
1000 Hilltop Circle, MD 21250
finin@cs.umbc.edu

Deborah L. McGuinness
Tetherless World Constellation
Rensselaer Polytechnic Institute
110 8th St., Troy NY 12180
dlm@cs.rpi.edu

ABSTRACT

Linked Data is a steadily growing presence on the Web. In Linked Data, the description of resources can be obtained incrementally by dereferencing the URIs of resources via the HTTP protocol. The use of owl:sameAs further enriches the Linked Data space by declaratively supporting distributed semantic data integration at the instance level. When consuming Linked Data, users should be careful when handling owl:sameAs: in that URIs linked by owl:sameAs may not be appropriate for simple aggregation, and that recursively exploring owl:sameAs may lead to considerable network overhead. In this work, we discuss and conduct an empirical pilot study on the usage of owl:sameAs in the Linked Data community. The results include initial quantitative measures of the usage of owl:sameAs. Based on observations of these results, we further discuss several strategies for dealing with owl:sameAs in Linked Data applications.

Keywords

Linked Data, owl:sameAs, experiment

1. INTRODUCTION

Linked Data [2] enables machines to surf the Semantic Web. By publishing data in RDF formats in accordance with common conventions, data publishers enable Linked Data applications to incrementally expand their knowledge about Semantic Web resources. In the body of Linked Data published thus far, `owl:sameAs` is increasingly used to provide declarative semantics for aggregating distributed data. That is, machines can merge resource descriptions if the resources described are linked with owl:sameAs.

The rising use of `owl:sameAs` can be observed in many important Linked Data datasets such as DBpedia[1], Freebase[2], GeoNames[3] and New York Times[4]. Examination of the 2009 Billion Triples Challenge dataset[5] further reveals some 6.5 million `owl:sameAs` statements. Instead of jumping into those huge Linked Data datasets at the Web scale, we have conducted a pilot study on the emerging usage of `owl:sameAs` in a subset of the Linked Data cloud.

Our study has three parts: (i) a review of the known issues with `owl:sameAs` reported by Web developers and researchers; (ii) the design of several methods and metrics for quantitatively measuring `owl:sameAs` usage and potentially discovering new uses of `owl:sameAs`; the methods and metrics are tested in a small subset of linked data; and (iii) a discussion of empirical strategies for dealing with `owl:sameAs` especially in linked data consumption. Our work so far focuses on the use of `owl:sameAs` in practice, as opposed to the official formal semantics of owl:sameAs. Interested readers may look into a parallel work by Halpin and Hayes[4] which qualitatively discusses various uses of owl:sameAs.

2. KNOWN ISSUES WITH OWL:SAMEAS

The `owl:sameAs` property is part of the Web Ontology Language (OWL) ontology[1]. It is frequently used to support Linked Data integration via declaratively interconnecting "equivalent" resources across distributed datasets. However, more researchers and developers have found the use of `owl:sameAs` does not always conform to its formal semantics defined in OWL. In the rest of this section, we review several observations along this line.

2.1 From rdfs:seeAlso to owl:sameAs

Prior to the rise of `owl:sameAs`, the `rdfs:seeAlso` property was heavily used in linking Friend of a Friend (FOAF) data: it links from one FOAF document to another in which additional descriptions about the resource can be found. Typically, the property `rdfs:seeAlso` is used with an instance of `owl:InverseFunctionalProperty`, e.g. `foaf:mbox_sha1sum`, in resource description, so that user can use identity information to find the matching resource in the remote FOAF document. More recently, `owl:sameAs` has been widely used in linked data datasets, such as DBpedia, and it provides an alternative way to refer to an external equivalent resource: the dereferenceable HTTP URI plays the role of rdfs:seeAlso and the URI itself can uniquely identify the

[1] `http://dbpedia.org/` - DBpedia
[2] `http://rdf.freebase.com/` - Freebase
[3] `http://sws.geonames.org/` - GeoNames
[4] `http://data.nytimes.com/` - New York Times
[5] `http://vmlion25.deri.ie/index.html`

Web Science Conf. 2010, April 26-27, 2010, Raleigh, NC, USA.

matching resource in the remote document. Moreover, the usage of `owl:sameAs` is no longer limited to the FOAF domain. The following is a fragment from Tim Berners-Lee's FOAF profile [6] which illustrates the usage of `owl:sameAs` as well as `rdfs:seeAlso`.

```
<con:Male
  rdf:about="http://www.w3.org/People/Berners-Lee/card#i">
  <owl:sameAs rdf:resource="http://identi.ca/user/45563"/>
  <foaf:knows rdf:resource="#dj"/>
</con:Male>
<foaf:Person rdf:about="#dj">
  <rdfs:seeAlso
     rdf:resource="http://www.grorg.org/dean/foaf.rdf"/>
  <foaf:mbox_sha1sum>6de4ff27ef927b9ba21ccc88257e41a2d7e7d293</
   foaf:mbox_sha1sum>
  <foaf:name>Dean Jackson</foaf:name>
</foaf:Person>
......
```

2.2 owl:sameAs is Not Symmetric

In 2007, Vatant [8] suggested that `owl:sameAs` is not a symmetric property and an agreed `owl:sameAs` relation should be supported reciprocally by both owners of the resources connected by owl:sameAs. The following example (copied from the author's original post) demonstrates this: the agreement about the equivalence of two resources, namely $a : foo$ and $b : bar$, can be confirmed if their owners, (a) and (b) respectively, have asserted `owl:sameAs` statements.

```
(a) asserts "a:foo owl:sameAs b:bar"
(b) asserts "b:bar owl:sameAs a:foo".
```

A more detailed account of the use of owl:sameAs in practice can be found in [4]. For example, even two URIs linked by `owl:sameAs` do refer to the same thing, their descriptions often should not be simply aggregated by merging graphs.

2.3 owl:sameAs Closure

In 2007, Passant [7] has shown that `owl:sameAs` can be used to support mashing up a person's information from different social networks such as Flickr. While most of the early instances of the `owl:sameAs` property connect a described resource to a regular web URI (see example above on Tim Berners-Lee's FOAF profile), more recent uses of `owl:sameAs` connect resources from linked data sources. The following is an example of a fragment of an `owl:sameAs` network. This example shows the New York Times is adding links to a Geonames URI that is not available in DBpedia and Freebase adding new `owl:sameAs` links to many DBpedia resources which are either non-existent or simply a DBpedia "redirection".

```
dbpedia:Virginia
  owl:sameAs nyt:N53394720474045997421;
  owl:sameAs freebase:guid.9202a8c04000641f8000003f833.

freebase:guid.9202a8c04000641f800000000003f833
  owl:sameAs dbpedia:Virginia_(state);
  owl:sameAs dbpedia:Commonwealth_of_Virginia;
  owl:sameAs dbpedia:Rest_of_Virginia;
  owl:sameAs dbpedia:Virginia;
  owl:sameAs dbpedia:Climate_in_virginia .

nyt:N53394720474045997421
  owl:sameAs nyt:virginia_geo ;
  owl:sameAs dbpedia:Virginia;
  owl:sameAs <http://sws.geonames.org/6254928/> .
```

[6] `http://www.w3.org/People/Berners-Lee/card`, the foaf namespace and rdfs namespace have been added to improve readability.

2.4 Simple Aggregation May Cause Errors

The equivalence relationship represented by `owl:sameAs` is often context-dependent, and is accurate only in the context of one application [5]. Therefore, the use of `owl:sameAs` in Linked Data may conflate context-dependent descriptions provided in different data sources. As shown in the following example, Li Ding has two FOAF profiles. The one hosted at Stanford University was accurate when it was published several years ago, but some facts have changed since then. A more recent FOAF profile indicates that he is now working at RPI and holds a job title of "Research Scientist". Each profile uses a unique URI to identify the person Li Ding, and it is reasonable to declare the two URIs are referring to the same person. However, if we connect the two URIs using owl:sameAs, an OWL reasoner can infer, on integrating the two datasets, that Li Ding holds the position of "Research Scientist" at Stanford University, which has never been the case.

```
<http://ksl.stanford.edu/people/ding/foaf.rdf#dingli>
  foaf:schoolHomepage <http://www.stanford.edu> .

<http://www.cs.rpi.edu/~dingl/foaf.rdf#me>
  cv:job_position "Research Scientist" .
```

Similar issues were raised about an earlier version of the New York Times dataset, where the `cc:license` property could be wrongly propagated to DBpedia's resource description via `owl:sameAs` inference [3].

This concern can further lead to conflicting statements from different sources. Consider, for example, the population of Warsaw, the capital of the county Poland. As shown in the following example, two different numbers were obtained from DBpedia and Geonames respectively. Each value could be true in a certain context; however, in answering a simple question "what is the population of Warsaw", web users are expecting the result to be just one number rather than a set of alternatives.

```
dbpedia:Warsaw
    dbpprop:populationTotal "1 709 781"@en.

<http://sws.geonames.org/756135/>
   Geonamesprop:population "1702139".
```

Moreover, a reasonable ontology might define the population property for places to be a sub-property of `owl:FunctionalProperty`. While this may be good modeling in theory, in practice it will lead to contradictions in cases like this. The current DBpedia ontology does not have domain and range constraints or cardinality restrictions for just this reason. This remains a conflict between modeling theory and practice that waits for solutions. This issue can also be viewed as being related to the challenge of maintaining provenance for independently generated objects that later are connected via sameas. McCusker and McGuinness [6] discuss this issue in the context of sameas usage in biomedical settings.

3. MEASURING OWL:SAMEAS USAGE

In order to better understand the reality of how `owl:sameAs` is being used, we carried out a simple empirical study on a dataset generated from a small set of seed URIs. We

were, in particular, interested in collecting and analyzing naturally occurring `owl:sameAs` networks. An `owl:sameAs` network is a set of URIs interconnected by `owl:sameAs` relations. In what follows, we first explain how we built a small evaluation dataset, go through the metrics we defined such networks and make some observations about the data.

3.1 Evaluation Dataset

All of the seeds were selected from the New York Times (NYT) dataset, which is a popular and carefully-created source of linked data containing a significant number of `owl:sameAs` statements. To keep our study manageable, we selected 100 seed URIs for each of the three distinctive categories from the NYT corpus: people, locations and organizations. For each seed URI, we performed a Web crawl by dereferencing the seed URI and the URIs (transitively) linked from the seed URI by `owl:sameAs` statements. The crawling process yielded a total of 4352 URIs from 300 networks in the three entity categories. There are only 3533 (81%) dereferenceable URIs contributing 117361 triples. Since almost every NYT corpus URI links to and back from DBpedia, we can see that this dataset is reflects the usage of `owl:sameAs` in DBpedia related linked data. We kept out dataset small to help us focus on out quantitative metrics for measuring usage of `owl:\-sameAs`.

3.2 Scale of the owl:sameAs Network

How large is an `owl: sameAs` network, where do the nodes come from, and which part is useful? The size of a network can be measured in terms of the number of nodes it contains. The source of the URIs can be identified by the hostname of the website hosting the URIs by analyzing the namespace of the URI. The usefulness of a URI can be predicted based on two factors: whether it is dereferenceable and whether it carries useful descriptions. Based on the above observations, we designed the following metrics:

- **size of network** is measured by the number of unique URIs in `owl:sameAs` network
- **dereferenceable portion of network**(d) is measured by the number of dereferenceable URIs in the `owl:sameAs` network.
- **useful portion of network**(u) is measured by the number of URIs in the `owl:sameAs` network which has been described by more one triple.

Figure 1 was generated by computing the above metrics on our evaluation dataset. We make five readily apparent observations. (i) On average, an `owl:sameAs` network involves 10 to 20 URIs, contributed by DBpedia, Freebase, and the New York Times. (ii) Location related URIs typically have a larger network. (iii) DBpedia is consistently the major contributor to URIs in these networks. (iv) There exist some non-dereferenceable URIs in all categories, and these URIs are either in DBpedia or other unlisted websites. (v) DBpedia has contributed many URIs described by just one triple as they correspond to Wikipedia "redirection" links, and New York Times similarly redirects human-readable URIs to their permanent counterparts which are distinguished by numbers.

3.3 Individual Contributions

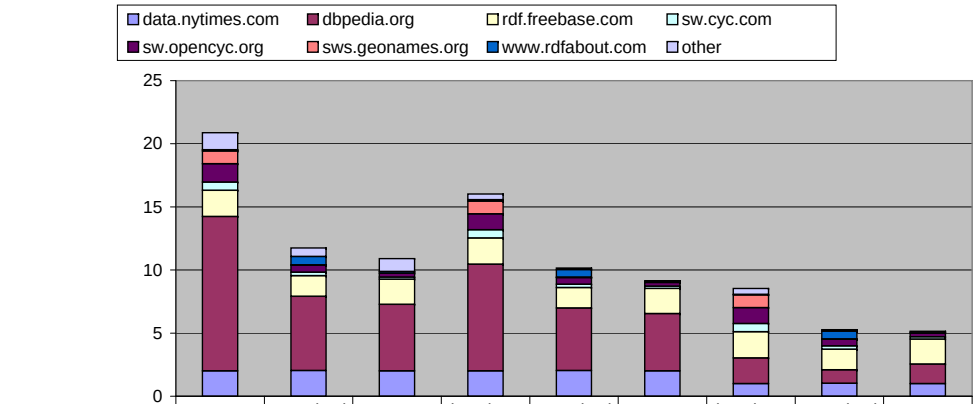
How much information does each URI contribute? Does it contribute new information, confirm existing information or contradict existing information? In order to answer these questions, content analysis is needed. Instead of fully relying on human intelligence, we designed several metrics that can be automatically computed as follows (the metrics can be used to guide in-depth analysis). We restricted our study to analyzing properties and literals, as they are the primary raw information carriers while the other URIs in resource description are more responsible for capturing structures. To keep this study simple, we leave analysis of other URI usage for future work.

- **Information richness** can be measured by simply counting the number of triples used by the resource description of each URI. It provides a rough estimate of the amount of information contributed by individual URIs.
- **Property usage** can be measured by simply counting the reused and unique properties in the resource descriptions of all URIs. It helps users understand the perspectives of resource descriptions of individual URIs. Future work may include analysis using the local name of properties or even the list of words extracted from the local name.
- **Literal usage** can be measured by counting the reused and unique literal strings in the resource descriptions of all URIs. It helps users understand the actual information conveyed by a resource description.

Based on analysis of information richness, we generated Figure 2 and observe the following (i) among the 4352 discovered URIs, only a few URIs are contained in a lot of triples; (ii) over a quarter of all URIs are only contained in one triple each, and most of them are found carrying a "redirection" meaning; and (iii) 90% of the triples contain 20% of the URIs, each of which is containe in more than 25 triples.

By analyzing property usage, we found that (i) URIs from the same source use a common set of properties; (ii) URIs from different sources seldom reuse properties; (iii) a few concepts from several domain independent popular ontologies were reused, e.g. `rdfs:label`, `geo:lat`, `foaf:homepage` and `geonames:population`; and (iv) some properties are seen frequently used in expressing contextual information, such as `cc:attributionName` in the Freebase dataset and `nyt:first_use` in NYT dataset.

With literal usage analysis, we may partially tackle the computation of difference of RDF graphs. We compute a rough estimate of the "RDF diff" by counting reused and unique literals based on the intuition that literals in an RDF graph constitute the main body of informational knowledge since they can be read by the end users directly. Our analysis shows that (i) among the URIs from a source, only a few (typically just one) contribute a lot of informational triples, and the rest may only contribute very few triples. (ii) As shown in Figure 3, DBpedia and Freebase are the primary sources of literals (accounting for 83%) but they don't have many literals in common. (iii) Further manual analysis on the values shows potentially conflicting literal values, e.g. the postcode in GeoNames follows the five digit ZIP code standard but DBpedia serves zip codes using the extended ZIP+4 code.



	locations	organizations	people	locations (d)	organizations (d)	people (d)	locations (u)	organizations (u)	people (u)
other	1.36	0.67	1.02	0.46	0.13	0.07	0.46	0.13	0.07
www.rdfabout.com	0.1	0.67	0.14	0.1	0.6	0.07	0.04	0.6	0.07
sws.geonames.org	1.01	0.03	0	1.01	0.03	0	1.01	0.03	0
sw.opencyc.org	1.45	0.55	0.3	1.26	0.53	0.3	1.26	0.53	0.3
sw.cyc.com	0.65	0.28	0.17	0.65	0.26	0.17	0.65	0.26	0.17
rdf.freebase.com	2.08	1.63	1.99	2.08	1.63	1.99	2.08	1.63	1.99
dbpedia.org	12.23	5.87	5.28	8.46	4.95	4.55	2.03	1.07	1.54
data.nytimes.com	2	2.04	2	2	2.03	2	1	1.02	1

Figure 1: Number of URIs (colored by source) per owl:sameAs network

4. DISCUSSIONS

Our experimental results have led us to identify several issues involving the `owl:sameAs` property as it is used in practice in a linked data context. These include how best to manage `owl:sameAs` assertions from "third parties", problems in merging assertions from sources with different contexts, and the need to explore an operational semantics distinct from the strict logical meaning provided by OWL.

4.1 Third Party's Contribution

A further question on `owl:sameAs` publishing is that of how to deal with third-party asserted `owl:sameAs` relations. The URI owners (who own the namespaces of URIs and contribute the official descriptions associated with the URIs) may not be interested in making `owl:sameAs` assertions; therefore, third-party asserted `owl:sameAs` relations could be used to facilitate linked data integration. Indeed, *sameas.org* is playing this role and has collected millions of third-party asserted `owl:sameAs` relations. It would be good to both promote reciprocal `owl:sameAs` confirmation mechanisms and develop effective trust mechanisms to assure the quality of `owl:sameAs` relations.

4.2 Projection based Partial Equivalence

Many `owl:sameAs` statements are asserted due to the equivalence of the primary feature of resource description, e.g. the URIs of FOAF profiles of a person may be linked just because they refer to the same person even if the URIs refer the person at different ages. The odd mashup on job-title in previous section is a good example for why the URIs in different FOAF profiles are not fully equivalent. Therefore, the empirical usage of `owl:sameAs` only captures the equivalence semantics on the projection of the URI on social entity dimension (removing the time and space dimensions). In this way, `owl:sameAs` is used to indicate partial equivalence between two different URIs, which should not be considered as full equivalence.

Knowing the dimensions covered by a URI and the dimensions covered by a property, it is possible to conduct better data integration using `owl:sameAs`. For example, since we know a URI of a person provides a temporal-spatial identity, descriptions using time-sensitive properties, e.g. age, height and workplace, should not be aggregated, while time-insensitive properties, such as eye color and social security number, may be aggregated in most cases.

4.3 Operational Semantics for Linked Data Consumption

We can find many examples where the the likely meaning of an `owl:sameAs` assertion in Linked Data is intended to be the official semantics as defined by OWL. Nonetheless, we cannot assume that it is never used with the intended semantics of absolute identity in mind. Since suitable alternatives to `owl:sameAs` do not exist (or are rarely used

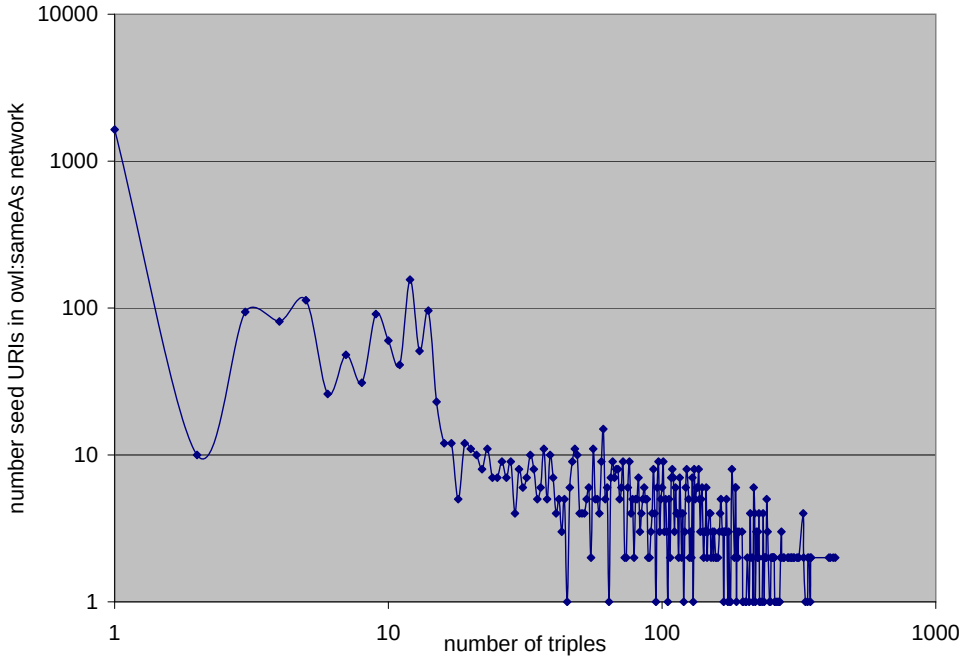


Figure 2: log-log plot of the count of seed URIs contributing exactly n triples

in practice), a Linked Data application is forced to make a choice with respect to each `owl:sameAs` link it encounters. In order to keep the information it gathers as consistent as possible. We propose several components of a general strategy for integrating and fusing information from the URIs in an `owl:sameAs` network.

- **Complementary descriptions**: if the associated descriptions of the URIs linked by owl:sameAs are orthogonal (taking into account the transitivity of owl:sameAs), then they can safely be merged. In linked data consumption, this kind of URIs should be dereferenced to collect the most complete description of the resource. For example, New York Times and DBpedia are complementary: the former provides news-oriented resource descriptions while the latter focuses on general descriptions about the same resource.

- **Alternative descriptions**: if the associated descriptions of the linked URIs are asserting different values for the same property, conflicts may occur when users expect a unique value from the property. For example, users expect at most one value for a population property. Moreover, when merging resource descriptions using correlated properties, conflicts may occur when, e.g., merging `foaf:firstName` and `foaf:surname` from different sources. One good example of alternative semantics can be found in proofs: users expect that one conclusion is exclusively justified by exactly one proof. In linked data consumption, only one of the fully alternative URI should be dereferenced, and the rest should be discarded.

- **Reconcilable descriptions**: if the associated descriptions of the linked URIs are neither fully orthogonal nor fully alternative, users may have more options. They can simply filter the portion of conflicting descriptions in an application-specific way, by taking the context of the descriptions into account. For example, an application mashing us description of a person may selectively aggregate eye color information but not the age information. In linked data consumption, all such URIs should be dereferenced, but only part of descriptions are going to be aggregated.

- **Redundant descriptions**: if the associated descriptions of the linked URIs forms a subset (or implication) relations, only the URI with broader coverage needs to be dereferenced. For example, if we know the information provided by source A was essentially part of the information provided by another source B, we can skip any URI derferencing operations at A if the corresponding URI at B has already found and dereferenced. This component can be used to deal with linked data generated from the same source, e.g. DBLP.

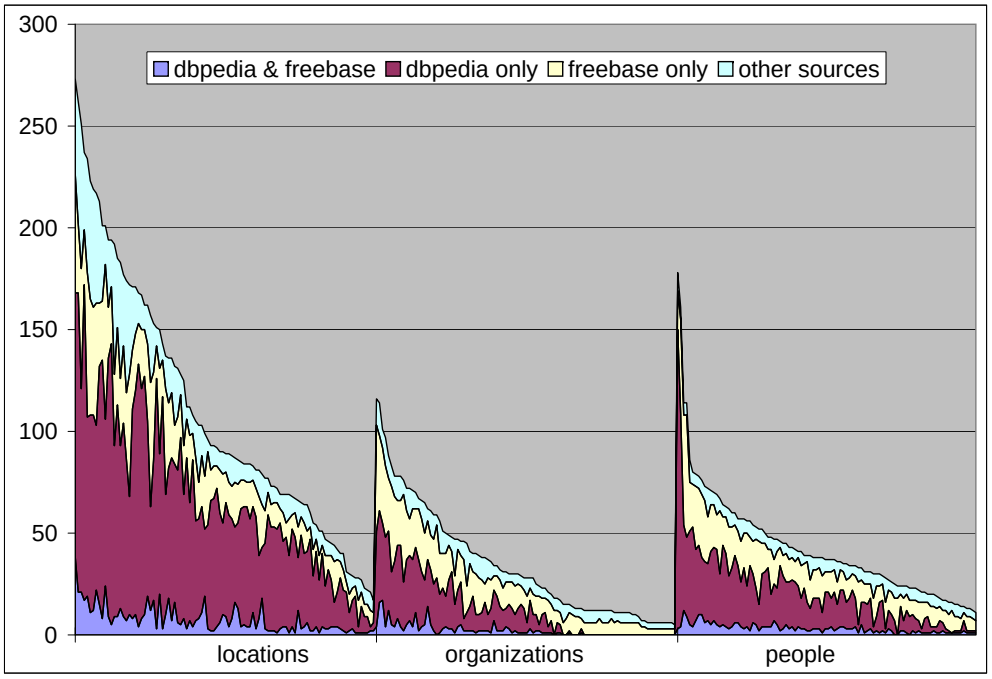


Figure 3: Sources of literals in RDF graph dereferenced from URIs in the owl:sameAs network of a seed URI

5. CONCLUSION

This brief empirical study shows some interesting directions related to `owl:sameAs`. It also suggests emerging operational semantics of `owl:sameAs` in linked data consumption. In particular, we have found that some URIs can be integrated in exclusive manner, i.e. such that only one should be chosen, as opposed to truly indistinguishable resources whose descriptions should be merged. By selecting the right interpretation of an `owl:sameAs` relation, we can reduce the potential overhead in retrieving and storing associated RDF descriptions. Given the URI of a resource described in Linked Data, we have the option of either dereferencing and merging all equivalent resources, based on `owl:sameAs` statements, or of picking and choosing alternatives based on our knowledge about the distributed data sources.

6. ACKNOWLEDGEMENT

This work is funded in part by a grant from DARPA's Transformational Convergence Technology Office, a gift from Microsoft Corporation and NSF award IIS-0326460.

7. REFERENCES

[1] S. Bechhofer, F. van Harmelen, J. Hendler, I. Horrocks, D. McGuinness, P. Patel-Schneider, and L. A. Stein. OWL Web Ontology Language Reference. W3C Recommendation, February 2004. `www.w3.org/TR/owl-ref`.

[2] T. Berners-Lee. Linked data. http://www.w3.org/DesignIssues/LinkedData.html, 2007.

[3] T. Berners-Lee. Linked data at the new york times: Exciting, but buggy. http://dowhatimean.net/2009/10/linked-data-at-the-new-york-times-exciting-but-buggy, 2009.

[4] H. Halpin and P. J. Hayes. When owl:sameAs isn't the same: An analysis of identity links on the semantic web. In *Proceedings of the 2010 International Workshop on Linked Data on the Web*, April 2010.

[5] A. Jaffri, H. Glaser, and I. Millard. Uri disambiguation in the context of linked data. In *Proceedings of the 1st International Workshop on Linked Data on the Web*, April 2008.

[6] J. McCusker and D. L. McGuinness. owl:sameas considered harmful to provenance. In *Proceedings of the ISCB Conference on Semantics in Healthcare and Life Sciences*, Feburary 2010.

[7] A. Passant. :me owl:sameas flickr:33669349@n00. In *Proceedings of the 1st International Workshop on Linked Data on the Web*, April 2008.

[8] B. Vatant. Using owl:sameas in linked data. http://blog.hubjects.com/2007/07/using-owlsameas-in-linked-data.html, 2007.