Assessing the Quality of owl:sameAs Links

Pierre-Henri Paris

Conservatoire National des Arts et Métiers,
CEDRIC, 292 rue saint martin, Paris, France,
pierre-henri.paris@upmc.fr

Abstract. *owl:sameAs* is one of the most important properties of the Linked Open Data domain. The property is used to indicate that two things are the same and when these two things come from two different datasets. But unfortunately, there is a gap between the purpose of this property, the way it was conceived and designed, and the way it is used in the wild. This deficiency is mainly due to the misuse of *owl:sameAs* because its strict semantics is not always taken into account or well understood and there may be no clear alternative. As a result, inaccurate data can be either crawled or inferred from incorrect (or at least questionable) links. Depending on the need for high quality data, the impact on the end users of the data can be considerable. This is why we propose to study how erroneous links can influence data, how to repair such cases, or even better, how to prevent the production of these links. We want to assess wrong links impact (e.g. at which points they decrease overall quality) and how they can be addressed.

Keywords: Linked Open Data, identity, SameAs, OWL

1 Introduction

Linked Open Data[1] (LOD) is an initiative proposed by Tim Berners-Lee to publish data on the Web. Such datasets use standards from the Semantic Web (SW) like the Resource Description Framework[2] (RDF) to represent data. RDF is an oriented and labeled multi-graph model. LOD datasets are graphs and thus can be linked together through the addition of statements having one node in one dataset and another node in a different dataset. This is the linking part of Linked Open Data.

There are more and more LOD datasets published and linked together. Linking datasets is a powerful way to enable retrieving knowledge from other datasets. This enables the ability of crawl various datasets to discover new facts. Thanks to ontologies using OWL [16], it is even possible to use reasoners based on, e.g., Description Logics (DL), to infer new things from data.

There are many ways to link two datasets. Many properties can be used to do so, since virtually any object property[3] can be used. One of the most employed is

[1] http://5stardata.info/en/

[2] https://www.w3.org/TR/2004/REC-rdf-primer-20040210/

[3] https://www.w3.org/TR/owl-guide/

the famous *owl:sameAs* property from the OWL ontology. To create a link with *owl:sameAs* between two instances is stating that these two instances are the same. It is an identity link. For example, let us say that we have two instances *ds1:Paris* and *ds2:CityOfParis* of Paris, the French capital. Each one of these instances is in one dataset (*ds1* and *ds2* here). It is possible to link *ds1* and *ds2* by simply adding this fact: *owl:sameAs(ds1:Paris, ds2:CityOfParis)*.

Because *owl:sameAs* has a very strict semantics, creating such links among instances has consequences. *owl:sameAs* semantics is based upon Leibniz's principle of indiscernibility. Therefore if two things are identical they must share the same values for the same properties. If in *ds1* the population of Paris is not filled out but is given in *ds2*, then the population can be used in *ds1*. More formally, the indiscernibility of identicals is: $owl{:}sameAs(a, b) \rightarrow (p(a, o) \rightarrow p(b, o))$. This is how one can discover new knowledge through the use of *owl:sameAs* links, by either retrieving or inferring new data. An erroneous link can lead to inaccurate inferred data and consequently it is very important to have high quality *owl:sameAs* links in order to maintain the overall quality of the datasets. As stated in Ding et al. [6] and Halpin et al. [12], there are many misuses of *owl:sameAs* in current LOD, which undermines the quality of the data it provides.

2 State of the Art

There are many facets to the interlinking problem. We want to underline four of them in this section.

2.1 Instance matching

The term instance matching refers to the problem of finding equivalent resources. As stated in Hogan et al. [15], the search for identity links among instances of the *LOD* has several names in literature such as *data linking* [10], *data reconciliation* [25], *record linkage* [21], *duplicate identification* [7], *object consolidation* [14], *instance matching* [3], *link discovery*, or *Co-reference resolution*. Those approaches are historically the first ones to emerge. The goal is to produce links between a source dataset and a target dataset. For each potential link a similarity score is produced and if the score is above some threshold, the link is validated. Ferraram et al. [8] published a complete survey and more recently Achichi et al. [1] and Nentwig et al. [20] propose complementary surveys.

2.2 Knowledge enrichment

Because several approaches used to match instances benefit from semantics features like inverse functional properties or cardinality restrictions, knowledge enrichment approaches are tangential to our domain of concern. For example, if one can find that a property is inverse functional, then any subjects related to the same object by this property have to be the same. More formally:

$InverseFunctional(P) \land P(a,c) \land P(b,c) \rightarrow owl{:}sameAs(a,b)$. Thus, the addition of new knowledge to the *TBox*[4] of an ontology might, hypothetically, lead to finding novel *owl:sameAs* links. Völker and Niepert [27] propose the induction of a schema (i.e. an ontology) for a given KB by using statistics. Association rules are mined and then translated into OWL2 axioms, but some interesting features cannot be mined, e.g. inverse properties, cardinal restrictions or property disjointness. Töpper et al. [26] also use statistical methods but from the Inductive Logical Programming (ILP) field, where only property domain, property range and class disjointness are computed. With AMIE [9], the authors use an assumption called the partial completeness assumption (PCL) and calculate scored rules under this assumption. PCL assumes that if a subject-predicate pair is present in the KB, then all possible objects for that pair are in the dataset. This method does not take into account any existing ontological knowledge nor use reasoning capabilities of Description Logics (DL). The approach of d'Amato et al. [4] is able to use ontological knowledge to produce rules.

2.3 Identity crisis

Since several works, like Halpin et al. [12] or Ding et al. [6], raised the misuse of identity links, proposals have been made to circumvent this problem. Halpin et al. [12] propose to use weaker versions of *owl:sameAs* (e.g. from *SKOS* vocabulary) but with a loss of inference capabilities. There are also proposals for new properties to represent identity relations by McCusker and McGuinness [19] and Halpin et al. [12], where *owl:sameAs* is a sub property of a more specific and more relaxed list of (semi-)identity properties. [12] by using a new ontology where *owl:sameAs* is a sub property of more specific and more relaxed link of (semi-)identity links. In this hierarchy each property has a combination of reflexivity, symmetry and/or transitivity. McCusker and McGuinness [19] propose a local and domain-specific approach where one can specialize *owl:sameAs* property in a particular domain (e.g. using a *biomedidentity:sameAsBioSource* property in the biology domain) but link usability thus becomes only local. At the intersection of link invalidation and contextual identity links, De Melo [5] propose to use a property to assert proven identity links on the basis that *owl:sameAs* might contain erroneous links. That is if an *owl:sameAs* link between a and b is proven to be right then one can have a *lvont:strictlySameAs* link between a and b. In Halpin et al. [13], authors propose to manage the context of identity links through the addition of a formal context. Several ideas are proposed, but none of them has been widely adopted. Beek et al. [2] propose to manage the context as a set of properties. Therefore two things are equal if they share the same property-value pairs where properties are defined by the context. Idrissou et al. [17] extends the proposition of Beek et al. [2] by adding operators other than the intersection between the sets of properties representing different contexts. Raad et al. [24], also based on Beek et al. [2] work, propose an algorithm to compute those contexts, i.e. contexts based on sets of properties where identity holds.

[4] `https://www.lesliesikos.com/tbox/`

2.4 Identity link assessment

Identity link assessment approaches consist of checking if a link is true or false. It does not create any link but does evaluate existing ones. Guéret et al. [11] propose to use classical network measures to assess existing links. De Melo [5] propose an approach using the unique name assumption within datasets (i.e. an instance has one name in a dataset) to spot sets of instances linked by *owl:sameAs* where at least one link is presumed to be wrong. Next, a linear programming algorithm is used to compute the wrong link. In Papaleo et al. [22], the authors propose a logical approach to detect such wrong statements. The algorithm tries to detect logical conflicts by using semantics features like functional properties in small sub-graph containing the two involved instances to assess. Thus, this approach strongly relies on semantics. Paulheim [23] propose to use data mining methods. First, links are represented in an embedded space. Second, an outlier detection algorithm is used to detect links that may be erroneous.

3 Problem Statement and Contributions

As we have seen in Section 1, a wrong *owl:sameAs* link can lead to inferring wrong data and therefore reduce the overall quality of datasets. As stated by Halpin et al. [12] there is an identity crisis in the sense that the *owl:sameAs* strict semantics is not always respected when used. To state that two things are the same is a very strong statement in LOD domain, with huge consequences. It is common to find instances that are nearly the same but not quite. For example, the city of Paris is **geographically the same** as the administrative department of Paris (it is an administrative subdivision). But in a **legal** context they are **two completely different things**. Another example is a glass of water belonging to a set of glasses. All glasses look the same and in most contexts they certainly are interchangeable but two glasses of this set are two physically different objects. What is identical or not is a philosophical question.

So there is confusion regarding how the *owl:sameAs* property has been defined and how it has been used. *owl:sameAs* has been created to be used only when things are really the same, **in all contexts**. But in the wild it is used in a more relaxed context, leading users to infer inaccurate facts.

There is a need to know how incorrect *owl:sameAs* links decrease the overall quality, and how frequent such incorrect links are. Also it is important to investigate the use of semantic features among LOD datasets. Because *owl:sameAs* links computation also relies on semantics features (e.g. functional properties, maximum cardinality, etc.).

Our goal is to identify quality defects related to data interlinking, then to assess them and finally, propose a way to correct them. Some questions that this work proposes to address include: How to evaluate a *owl:sameAs* link? What is the impact of erroneous links of this type on datasets? How to measure this impact and how to reduce the negative effect on dataset quality?

4 Research Methodology and Approach

For now, the approach we consider is the following, depending on preliminary results:

1. The first step consists in identifying quality defects due to data interlinking. As pointed out by several authors, *owl:sameAs* links are far from sure things. They can be incorrect and therefore produce incoherence in datasets. Thus, we propose to investigate several domains and try to find the nature of the underlying quality defects such as incompleteness (missing links) or incoherence. We will have to define efficient algorithms for data exploration and defect detection.

   But before before diving deeper into quality defects, we plan to reproduce independently some results from previous work (like in Halpin et al. [12]) and to gather useful and precise statistics about *owl:sameAs* links and semantics in general. By semantics we mean OWL properties or classes, because *owl:sameAs* links rely a lot on them. Moreover we need to establish how much this identity crisis is a problem.
2. For each type of defect, we need to associate a set of assessment methods and algorithms. The proposal of Papaleo et al. [22] relies mainly on semantic features like functional or inverse functional properties. Problems arise when semantic features are not sufficiently present in the vocabulary or the data. For example, according to the last version of DBpedia (October-2016), the ontology contains only 30 functional properties (1%) and zero inverse functional property (out of more than 2860 properties). 1.6M out of 4.6M instances (34%) use at least one of the functional properties. Conversely, other approaches rely on statistical, graph or mining techniques. A hybrid approach using semantics, statistics on data and vocabulary, network or graphs measures (etc.) could be used successfully.
3. The next logical step is to be able to correct detected quality defects. Such defects decrease data quality and hence, finding how to correct them can improve data quality. There are, at this time, not so many propositions to deal with existing defects, where the balance tends to lean towards the creation of links in the literature.
4. We want to implement the proposed solutions through a Web application and/or a Web service to support Data Interlinking evaluation and improvement. This is an important point because if the tooling is not good enough, we cannot expect people to adopt our work.
5. Finally, with the high volume of data composing the LOD cloud[5], it is not only a matter of Open Data but also a Big Data issue. Hence, the developed solutions should scale well.

[5] http://lod-cloud.net/

5 Preliminary or Intermediate Results

Until now, we have focused primarily on related work. Nevertheless, we have written two articles to study some ideas related to our problem.

One of the interesting aspects of linking datasets is that one can complete data in a dataset thanks to another one. Therefore, studying completeness might be interesting because to understand the impact of interlinking between two datasets one has to know more about completeness of datasets. Once two instances are connected thanks to an *owl:sameAs* link, one can retrieve properties from the first instance towards the second one (and conversely). As a consequence, completeness might be subject to variation and to be able to assess this variation before and after interlinking is important information. We want to be able to check how the *owl:sameAs* property, used to connect a local dataset to external datasets, can help improve the completeness of this given local dataset. Furthermore, even if it might not be obvious at the first look, completeness is also a very important part to interlink two datasets, i.e. to produce links. In fact, the more complete the datasets, the easier it is to find proof that two instances are the same. For example, with only the last name we can not tell if two persons are the same or not. But in addition with their first names and birth dates, we have more hints to decide whether they are the same or not. Hence, a method for assessing completeness can be very interesting given the two previous points. We published a study [18] to assess the completeness evolution of DBpedia. In this work we proposed to assess completeness by using a mining-based approach. Using a given set of instances ($\mathcal{I}$) of the same class, the first step consists in computing a set of properties that are frequently found among instances of $\mathcal{I}$. For each instance, the corresponding transaction is all its properties. Here we use frequent-itemset algorithm to do so. Once those set of properties (or transactions) found, we can compute the completeness $\mathcal{CP}$ of the set of instances $\mathcal{I}$:

$$\mathcal{CP}(\mathcal{I}) = \frac{1}{|\mathcal{T}|} \sum_{k=1}^{|\mathcal{T}|} \sum_{j=1}^{|\mathcal{MFP}|} \frac{\delta(\mathcal{P}(t_k), \hat{P}_j)}{|\mathcal{MFP}|} \tag{1}$$

such that $\mathcal{T}$ is the set of transactions associated with $\mathcal{I}$ (see second column of Table 2). $\mathcal{MFP}$ is the set of maximal frequent patterns computed in the first step, $\mathcal{P}(t_k)$ is the power set of the transaction t_k and δ is equal to one if the frequent pattern $\hat{P}_j$ is in $\mathcal{P}(t_k)$, zero otherwise.

Example 1. The completeness of the subset of instances in Table 1 regarding their transactions in Table 2 and $\mathcal{MFP} = \{\{\textit{director,musicComposer}\}, \{\textit{director, editing}\}\}$, would be:

$$\mathcal{CP}(\mathcal{I}') = (2 * (1/2) + (2/2))/3 = 0.67$$

This value corresponds to the completeness average value for the whole dataset regarding the inferred patterns in $\mathcal{MFP}$.

Table 1. A sample of DBpedia triples

Subject	Predicate	Object
The_Godfather	director	Francis_Ford_Coppola
The_Godfather	musicComposer	Nino_Rota
Goodfellas	director	Martin_Scorsese
Goodfellas	editing	Thelma_Schoonmaker
True_Lies	director	James_Cameron
True_Lies	editing	Conrad_Buff_IV
True_Lies	musicComposer	Brad_Fiedel

Table 2. Transactions corresponding to triples from Table 1

Resource	Transaction
The_Godfather	{director, musicComposer}
Goodfellas	{director, editing}
True_Lies	{director, editing, musicComposer}

We apply this approach on several versions of DBpedia, thus allowing us to study evolution when new data is added or changed. This permits us to have an efficient way to characterize an important aspect of data quality, i.e. completeness.

We are also investigating the way *owl:sameAs* links are created. We are currently writing a proposition that takes into account the structure of data (both TBox and ABox[6]) in a dataset. We believe that not all evidence (in one direction or another) have the same strength. To compare two instances from two distinct datasets we want to collect evidences. Evidence is either a proof of identity or a proof of difference between those two instances. If the two instances share a property and an object, it is therefore a proof they may be the same. Rather, if they share a property but the objects are different, this evidence will tend to prove that the instances are different. The strength of this proof depends on the weight of the property among instances of the same class, and it also depends on the discriminating power of the object(s). The higher the weight and discriminating power are, the stronger the evidence. This approach can help us determine whether an existing relationship has a justification.

6 Evaluation Plan

The rudimentary version of our evaluation plan is as follows:

1. In order to validate our work, we will first select several interlinked datasets from the LOD. Datasets will have to cover different domains, with different sizes and vocabularies to ensure a large number of situations to be tested. For the moment, DBpedia and Wikidata are good candidates to be used as a playground since both datasets are cross-domain, well known and well

[6] `https://www.lesliesikos.com/abox/`

interconnected. Datasets in a specific domain should also be part of our assessment. Therefore, we also plan to use life sciences datasets, as there has been an explosion in the number of interlinked datasets in this area. Datasets from other domains can be used later if necessary.

2. For each dataset, we will have to evaluate several dimensions of data quality (still to be chosen).
3. The third step is to apply our approach. We will assess the interlinked datasets and modify them based on the experience gained in previous steps (e.g., to help determine which links can be safely deleted, which are useful in one context but not in another, etc.). This step will depend in large part on our future findings, so it may be subject to change.
4. Finally, we will again compute the same data quality dimensions. At this stage, we will have assessed several dimensions of data quality before and after the application of our approach, and so we can compare the situation before and after, hoping that there will be an improvement.

7 Conclusions

Being able to identify inaccurate links is a first step towards improving the overall quality of the data, since the data producer could be notified at the time of publication. Moreover, since the correction of links is not sufficiently addressed in the literature, if we achieve this objective, improving existing data sets would be a significant breakthrough in the identity crisis of the LOD.

Acknowledgements We would like to thank the anonymous reviewers for their helpful comments. We also sincerely thank Aidan Hogan for his expert advice.

References

1. Achichi, M., Bellahsene, Z., Todorov, K.: A survey on web data linking. Revue des Sciences et Technologies de l'Information-Série ISI: Ingénierie des Systèmes d'Information (2016)
2. Beek, W., Schlobach, S., van Harmelen, F.: A contextualised semantics for owl: sameas. In: International Semantic Web Conference. pp. 405–419. Springer (2016)
3. Castano, S., Ferrara, A., Montanelli, S., Lorusso, D.: Instance matching for ontology population. In: Italian Symposium on Advanced Database Systems. pp. 121–132 (2008)
4. d'Amato, C., Staab, S., Tettamanzi, A.G., Minh, T.D., Gandon, F.: Ontology enrichment by discovering multi-relational association rules from ontological knowledge bases. In: Proceedings of the 31st Annual ACM Symposium on Applied Computing. pp. 333–338. ACM (2016)
5. De Melo, G.: Not quite the same: Identity constraints for the web of linked data. In: AAAI National Conference of the American Association for Artificial Intelligence (2013)
6. Ding, L., Shinavier, J., Finin, T., McGuinness, D.L.: owl: sameas and linked data: An empirical study (2010)

7. Elmagarmid, A.K., Ipeirotis, P.G., Verykios, V.S.: Duplicate record detection: A survey. IEEE Transactions on Knowledge and Data Engineering 19(1), 1–16 (2007)
8. Ferraram, A., Nikolov, A., Scharffe, F.: Data linking for the semantic web. Semantic Web: Ontology and Knowledge Base Enabled Tools, Services, and Applications 169, 326 (2013)
9. Galárraga, L., Teflioudi, C., Hose, K., Suchanek, F.M.: Fast rule mining in ontological knowledge bases with amie + +. The VLDB Journal 24(6), 707–730 (2015)
10. Giannopoulou, I., Saïs, F., Thomopoulos, R.: Linked data annotation and fusion driven by data quality evaluation. In: EGC French Speaking Conference on the Extraction and Management of Knowledge. pp. 257–262 (2015)
11. Guéret, C., Groth, P., Stadler, C., Lehmann, J.: Assessing linked data mappings using network measures. In: Extended Semantic Web Conference. pp. 87–102. Springer (2012)
12. Halpin, H., Hayes, P.J., McCusker, J.P., McGuinness, D.L., Thompson, H.S.: When owl: sameas isn't the same: An analysis of identity in linked data. In: International Semantic Web Conference. pp. 305–320. Springer (2010)
13. Halpin, H., Hayes, P.J., Thompson, H.S.: When owl: sameas isn't the same redux: towards a theory of identity, context, and inference on the semantic web. In: International and Interdisciplinary Conference on Modeling and Using Context. pp. 47–60. Springer (2015)
14. Hogan, A., Decker, S., Harth, A.: Performing object consolidation on the semantic web data graph (2007)
15. Hogan, A., Polleres, A., Umbrich, J., Zimmermann, A.: Some entities are more equal than others: statistical methods to consolidate linked data. In: 4th International Workshop on New Forms of Reasoning for the Semantic Web: Scalable and Dynamic (NeFoRS2010) (2010)
16. Horrocks, I., Kutz, O., Sattler, U.: The even more irresistible SROIQ. KR 6, 57–67 (2006)
17. Idrissou, A.K., Hoekstra, R., van Harmelen, F., Khalili, A., van den Besselaar, P.: Is my: sameas the same as your: sameas?: Lenticular lenses for context-specific identity. In: Proceedings of the Knowledge Capture Conference. p. 23. ACM (2017)
18. Issa, S., Paris, P.H., Hamdi, F.: Assessing the completeness evolution of dbpedia: A case study. In: International Conference on Conceptual Modeling. pp. 238–247. Springer (2017)
19. McCusker, J.P., McGuinness, D.L.: Towards identity in linked data. In: OWLED Proceedings of OWL Experiences and Directions Seventh Annual Workshop (2010)
20. Nentwig, M., Hartung, M., Ngonga Ngomo, A.C., Rahm, E.: A survey of current link discovery frameworks. Semantic Web 8(3), 419–436 (2017)
21. Newcombe, H., Kennedy, J., Axford, S., James, A.: Automatic linkage of vital records (1967)
22. Papaleo, L., Pernelle, N., Saïs, F., Dumont, C.: Logical detection of invalid sameas statements in RDF data. In: International Conference on Knowledge Engineering and Knowledge Management. pp. 373–384. Springer (2014)
23. Paulheim, H.: Identifying wrong links between datasets by multi-dimensional outlier detection. In: WoDOOM Third International Workshop on Debugging Ontologies and Ontology Mappings-WoDOOM14. pp. 27–38 (2014)
24. Raad, J., Pernelle, N., Saïs, F.: Detection of contextual identity links in a knowledge base. In: Proceedings of the Knowledge Capture Conference. p. 8. ACM (2017)
25. Saıs, F., Pernelle, N., Rousset, M.C.: Combining a logical and a numerical method for data reconciliation. Journal on Data Semantics 12(12), 66–94 (2009)

26. Töpper, G., Knuth, M., Sack, H.: Dbpedia ontology enrichment for inconsistency detection. In: Proceedings of the 8th International Conference on Semantic Systems. pp. 33–40. ACM (2012)
27. Völker, J., Niepert, M.: Statistical schema induction. In: Extended Semantic Web Conference. pp. 124–138. Springer (2011)