On the Impact of sameAs on Schema Matching

Joe Raad, Erman Acar, Stefan Schlobach

▶ **To cite this version:**

Joe Raad, Erman Acar, Stefan Schlobach. On the Impact of sameAs on Schema Matching. K-CAP '19: Knowledge Capture Conference, Nov 2019, Marina Del Rey CA USA, France. pp.77-84, 10.1145/3360901.3364442 . hal-03416932

HAL Id: hal-03416932
https://hal.science/hal-03416932

Submitted on 5 Nov 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

On the Impact of sameAs on Schema Matching

Joe Raad
Vrije Universiteit
Amsterdam, The Netherlands
j.raad@vu.nl

Erman Acar
Vrije Universiteit
Amsterdam, The Netherlands
erman.acar@vu.nl

Stefan Schlobach
Vrije Universiteit
Amsterdam, The Netherlands
k.s.schlobach@vu.nl

ABSTRACT

In a large and decentralised knowledge representation system such as the Web of Data, it is common for data sets to overlap. In the absence of a central naming authority, semantic heterogeneity is inevitable as such overlapping contents are described using different schemas. To overcome this problem, a number of solutions have automated the integration of these data sets by matching their schemas. In this work we focus on a specific category of these solutions, which relies on the concepts' extension for matching the schemas (i.e., instance-based methods). Rather than introducing a new approach for the task of schema matching, this work studies the effect of exploiting the semantics of `owl:sameAs` in such instance-based methods. For this empirical analysis, we investigate more than 900K concepts extracted from the Web, and make use of over 35B implicit identity assertions to study their impact. The experiments show that despite the growing doubts over their quality, exploiting `owl:sameAs` assertions extracted from the Web can improve instance-based schema matching techniques.

CCS CONCEPTS

• **Information systems → Entity resolution**; **Semantic web description languages**; • **Computing methodologies → Knowledge representation and reasoning**;

KEYWORDS

linked open data, schema matching, identity

ACM Reference Format:
Joe Raad, Erman Acar, and Stefan Schlobach. 2019. On the Impact of sameAs on Schema Matching. In *Proceedings of ACM (K-CAP'19).* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The historic claim of the Semantic Web has been to foster interoperability of data sets published according to its formal principles. On the instance level, reusing resource identifiers and explicitly stating their equivalences through `owl:sameAs` statements have helped creating a huge Web of Data, with hundreds of thousands of linked data sets [2]. Historically though, most of these data sets use different schemas to model their data; thus, making reuse difficult, if not impossible. Over the past two decades, the Semantic Web community has targeted a lot of efforts on the task of *schema matching*, the task of identifying whether two concepts across different schemas are related. Various approaches [5] have been developed to determine whether a concept in a source schema is meant to refer to the same class of objects as a concept in a target schema, or in some cases to a more specific or more abstract class of objects. A wide variety of concept matching techniques were explored, ranging from terminological methods comparing labels and descriptions, via structural and graph-theoretic methods to extensional ones (i.e., instance-based methods).

In this study, we focus on the last category of approaches, where the concepts' set of instances are compared for deciding whether an equivalence between these concepts exists or not. Rather than proposing new measures for deciding whether a pair of concepts should be matched or not, this work studies the impact of exploiting instance-level interlinks in such schema-matching methods. Although instance-level interlinks can refer to various types of semantic relations between instances (e.g. *rdfs:seeAlso*, *owl:differentFrom*), this work considers only equivalence relations found in the form of `owl:sameAs` statements. With this study, we aim at providing instance-based schema-matching designers with empirical evidences on the benefits and drawbacks of using external collections of instance-level interlinks (e.g., from the LOD Cloud) in their tasks. Such study is particularly important, as it follows a number of analyses showing that a number of these `owl:sameAs` links are actually erroneous [8, 10, 17]. This uncertainty regarding the quality of existing `owl:sameAs` links, along with various other factors such as the way identity, typing and subsumption relations are published in the Web, poses the following two research questions:

- Q1 Does the inclusion of instance-level interlinks enhance instance-based schema alignments? (*w* and *w/o* considering the transitive closure of the class subsumption relation.)
- Q2 Is there a correlation between the quality of the instance-level interlinks and the quality of the resulting schema alignments?

Here, the two variations of Q1 can also be put as understanding the contribution of inference (restricted to subsumption) in enhancing the schema alignments.

For providing empirical answers for these two main research questions, we investigate more than 1K matched concepts and 900K unmatched concepts extracted from the Web of Data. We make use of over 558 million identity statements (35 billion after transitive closure), over 3 billion typing, and 4 million subsumption statements. In particular, we leverage the availability of two important elements of infrastructure, the *LOD-a-lot* data set [6], which makes thousands of linked data sets efficiently storable and queryable as an HDT (Header, Dictionary, Triples) file, and the *sameAs.cc* identity-cloud

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
K-CAP'19, November 2019, Marina del Rey, California, USA
© 2019 Copyright held by the owner/author(s).
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

[1] that was recently published. The latter is a queryable addition to the *LOD-a-lot*, representing the identity closure over its available owl:sameAs statements.

The rest of the paper is structured as follows. Section 2 presents related works. Section 3 presents the preliminaries and the notation. Section 4 presents our experimental settings. Section 5 presents our conducted evaluation, and Section 6 concludes the paper.

2 RELATED WORK

Instance-based schema matching. In its 2013 edition, the 'Ontology Matching' book [5] reviewed around 100 schema-matching systems. It classifies 15 as systems exploiting solely instance-level information for matching schemas, and an additional 27 systems as ones combining both instance- and schema-level information for this task. While their specific techniques might completely differ, all instance-based systems share two essential ideas: 1) the semantics of a concept is better determined by its members, rather than by its annotations, 2) the more significant the overlap between the two concepts' members, the more related these concepts are. The differences between these systems lie in the way the overlap between the concepts' members is measured, by for instance using formal concept analysis (FCA) techniques [18], machine learning [4], or classical similarity measures such as the Jaccard index [3, 11]. Certain instance-based approaches have already exploited [3] and computed [15, 19] instance-level interlinks for enhancing the quality of their schema alignments. The technique adopted by Correndo et al. [3] is closely related to this study, as the authors exploited less than 40K owl:sameAs links for matching DBpedia concepts. However, the experimental settings of the study presented here is orders of magnitude larger in terms of the number of considered instances and exploited owl:sameAs statements.

Identity Links in the Web of Data. Whenever multiple instance identifiers are used to denote the same real-world entity, identity statements are needed to link the data and foster reuse. The most commonly used relation for interlinking instances on the Web is the owl:sameAs[1] predicate, introduced in 2004 as part of the Web Ontology Language (OWL). This relation denotes a strict notion of identity, with a statement of the form $\langle x, \texttt{owl:sameAs}, y\rangle$ indicating that every property attributed to x must also be attributed to y, and vice versa (i.e., indiscernibility of identicals). The wide adoption of owl:sameAs has led to the emergence of several identity services [1, 7, 14] that harvest the Web, and offer access to these millions of identity statements with their transitive closure. The most recent, and comprehensive one in terms of the number of covered owl:sameAs, is the *sameAs.cc* identity service [1], providing access to more than 558 million owl:sameAs extracted from the Web. In addition to the emergence of such services, the special status of owl:sameAs led to a number of studies investigating the quality of these identity links [8, 10, 12, 17]. These studies showed that owl:sameAs is indeed used incorrectly; some studies estimating that around 3% [10] or 4% [16] of these links are erroneous, whilst others estimating this number to be in the range of 20% [8]. An example of such identity misuse and its inadvertent implications after transitive closure, is the largest equivalence class in *sameAs.cc*. This class, which in theory should include instances referring to the same real-world entity, contains in practice around 177K instances referring to various countries, cities, people, and religions. In this context, a recent approach [17] tried to limit the effects of such erroneous statements by assigning an error degree to each of these 558M owl:sameAs statements in *sameAs.cc* which is based on the community structure of the owl:sameAs network. These error degrees will be used in this study as indicators of the quality of owl:sameAs links.

To the best of our knowledge, this is the first work which studies the impact of instance-level interlinks on the quality of schema alignments. Such study has a potential impact on both families of related works. On one hand, it provides instance-based schema approaches with empirical evidences on the impact of considering owl:sameAs links for enhancing their schema systems. On the other hand, with this study based on the largest publicly available collection of owl:sameAs, it shows to which extent identity links extracted from the Web can be trusted to be deployed in certain Linked Data tasks, or whether a filtering process is required.

3 BACKGROUND AND NOTATION

In this section, we give the preliminary background, and also introduce the notation. For a more exhaustive background, we refer the reader to [9].

A knowledge base (KB) $\mathcal{K}$ is a set of RDF triples in the form of $\langle i_1, p, i_2\rangle$, where i_1 is an instance in the *subject* position and i_2 is an instance in the *object* position, and p is a *property*; representing a relation between i_1 and i_2. The knowledge graph $\mathcal{G}_\mathcal{K}$ induced by $\mathcal{K}$, is a tuple (V, E) where nodes V refer to the set of all named instances, and edges $E \subseteq V \times V$ refer to the set of all properties appearing in $\mathcal{K}$. For simplicity, we drop the index $\mathcal{K}$ in $\mathcal{G}_\mathcal{K}$, and use $\mathcal{G}$ instead whenever it is clear from the context.

Given a knowledge graph $\mathcal{G}$, let $\mathcal{G}^*_\sim = (V_\sim, E_\sim)$ be a subgraph of $\mathcal{G}$ in which every (existing) edge is an identity link i.e., owl:sameAs. Now, observe that for every $v \in V_\sim$, it is the case that $(v, v) \in E_\sim$. Since such reflexive identity links are trivial for our aims, we filter them out. Hence, we call instead $\mathcal{G}_\sim = (V_\sim, E_\sim \setminus E_{\texttt{id}})$ an *identity network* where $E_{\texttt{id}} = \{(v, v) \mid v \in V_\sim\}$. By the transitivity of the identity, every connected node i in $\mathcal{G}_\sim$ belongs to the same *equivalence class* which we will denote by $[i]$, calculated by the transitive closure of owl:sameAs (denoted by owl:sameAs*). Hence, for all $(i_1, i_2) \in$ owl:sameAs*, $[i_1] = [i_2]$. Note that since an identity network can not have an isolated node, an equivalence class in this case can not be a singleton.

Concepts, intuitively, refer to sets that possibly have named instances as *members* i.e., $\langle i, \texttt{rdf:type}, C\rangle$ is what we refer to when we say "instance i is stated as a member of concept C" or similar in short "i is a member of C". By the (simple) *extension* of a concept C, denoted by $ext(C) := \{i \mid \langle i, \texttt{rdf:type}, C\rangle \in \mathcal{K}\}$, we mean the set of instances which are explicitly stated as members of C, or in short *explicit members of* C. Let $\mathcal{K}^*$ be the KB obtained by the transitive closure of subsumption relation (i.e., rdfs:subClassOf) on $\mathcal{K}$. By an extension of C w.r.t. subsumption relation, denoted by $ext_\sqsubseteq(C)$, we mean all the instances that is either in $ext(C)$ or

[1] http://www.w3.org/2002/07/owl#sameAs

derived through concept subsumption (i.e., *implicit members of* C), hence $ext_{\sqsubseteq}(C) := \{i \in ext(B) \mid \langle B, \texttt{rdfs:subClassOf}, C\rangle \in \mathcal{K}^*\} \cup ext(C)$. The extension C w.r.t. to the equivalence class $\sim$ is defined as $ext^{\sim}(C) := \{j \in [i] \mid i \in ext(C)\}$. And last, the extension of C defined w.r.t. both equivalence class and the subsumption relation is defined as their union i.e., $ext^{\sim}_{\sqsubseteq}(C) := ext_{\sqsubseteq}(C) \cup ext^{\sim}(C)$.

The set of all concepts we consider are those that appear in the object positions of an RDF triple $\langle i, \texttt{rdf:type}, C\rangle \in \mathcal{K}$ with i called an instance, and is denoted by calligraphic $\mathcal{C}$. By $\mathcal{C}(i)$, we denote the set of concepts whose i is a member. Similar to the aforementioned notions of extensions, $\mathcal{C}_{\sqsubseteq}(i)$, $\mathcal{C}^{\sim}(i)$, and $\mathcal{C}^{\sim}_{\sqsubseteq}(i)$ are the sets of concepts which contains i w.r.t. subsumption, equivalence class, and the union of those two, respectively.

4 EXPERIMENTS

In this study, we aim at empirically measuring the impact of exploiting a collection of instance-level interlinks from the Web, on the quality of instance-based schema alignments. In other words, whether the addition of `owl:sameAs` links increase the similarity of two (in fact) equivalent concepts' extensions, without increasing the similarity of two non-equivalent ones. In practice, the exact impact of including `owl:sameAs` links will vary depending on the type of techniques used for measuring the similarity between the concepts' instance sets. For instance, FCA techniques might be more impacted by the inclusion of `owl:sameAs` links than machine learning techniques. In order to observe this impact independently from the type of technique deployed, we rely in this study on the simple Jaccard index for measuring the concepts' instance set similarity.

4.1 Jaccard Index with Equivalence Classes

The Jaccard index, denoted as J, is a commonly used measure to score the similarity between two sets [13] by ratio of their intersection over their union:

$$J(A, B) := \frac{|A \cap B|}{|A \cup B|}$$

where A and B are two sets. This index yields a value between 0 and 1, in which the higher the similarity of two sets is, the greater the Jaccard index.

Example 4.1. Given two concepts C_1 and C_2, with $ext(C_1) = \{i_1, i_2, i_3, i_4\}$, and $ext(C_2) = \{i_1, i_2, i_5\}$. With $ext(C_1) \cap ext(C_2) = \{i_1, i_2\}$ and $ext(C_1) \cup ext(C_2) = \{i_1, i_2, i_3, i_4, i_5\}$, the resulting $J(ext(C_1), ext(C_2))$ yields a value of 0.4.

Equivalence classes can provide further information about the instances of two sets of consideration. This additional information might result in either a positive or negative variation of the Jaccard index. Below we present these possible scenarios.

Scenario 1. *Equivalence classes increase Jaccard index.*

Let's assume the presence of an identity link between the instances i_3 and i_4 from the previous example, i.e., $\langle i_3, \texttt{owl:sameAs}^*, i_4\rangle$, hence both i_3 and i_4 belong to the same equivalence class $[i]$. In this scenario, replacing all instances that belong to the same equivalence class with a unique identifier $[i]^{ID}$ results in $ext^{\sim}(C_1) \cup ext^{\sim}(C_2) = \{i_1, i_2, [i]^{ID}, i_5\}$. With the decrease of their union size, while their intersection stays invariant, $J(ext^{\sim}(C_1), ext^{\sim}(C_2))$ increases to 0.5.

Another case where the Jaccard index increases is the presence of an identity link between instances from different instance sets, e.g., $\langle i_3, \texttt{owl:sameAs}^*, i_5\rangle$. In such scenario, $|ext^{\sim}(C_1) \cap ext^{\sim}(C_2)|$ increases and $|ext^{\sim}(C_1) \cup ext^{\sim}(C_2)|$ decreases, resulting in a higher increase of $J(ext^{\sim}(C_1), ext^{\sim}(C_2))$ to 0.75.

Scenario 2. *Equivalence classes decrease Jaccard index.*

Assuming the case from the previous example where i_1 and i_2 belong to the same equivalence class, $J(ext^{\sim}(C_1), ext^{\sim}(C_2))$ decreases to 0.25. In general, this is the case when equivalence classes apply mostly on the intersection set only. Indeed, since intersection is a subset of the union, same-size shrinkage on both sets has a higher impact on the size of the intersection, which results in an overall decrease on the Jaccard index.

Numerous cases in which the size of the intersection (union) of two instance sets increases (decreases) (i.e., Scenario 1) does not readily imply a positive impact of `owl:sameAs` on schema matching, since the Jaccard index of non-equivalent concepts might also increase. This is in strong connection to our first research question Q1 (which we shall give an empirical answer in upcoming sections). To settle this, we next investigate whether taking equivalence classes into account will increase the overlap of extensions for the correct mappings, and not for the incorrect ones.

Table 1: Statistics of the *LOD-a-lot* data set

# triples	28,362,198,927
# rdf:type	3,321,354,308
# owl:sameAs	558,943,116
# equivalence classes	48,999,148
# rdfs:subClassOf	4,461,717
# owl:equivalentClass	1,051,979
$\lvert C\rvert$	833,232
$\lvert C_{\sqsubseteq}\rvert$	976,674

4.2 Data sets & Implementation

In this section, we describe the data sets and the technologies deployed in this study. Table 1 summarises the main statistics of the data set described in this section.

Knowledge Base. We use the *LOD-a-lot* data set [6] as our knowledge base. This data set contains 28.3B triples collected from the 2015 *LOD Laundromat* crawl [2] of over 650K data documents from the Web. It is exposed in a single HDT file[2] that is 524GB in size, and is publicly accessible via an LDF (Linked Data Fragments) interface[3].

Identity Network & Equivalence Classes. We use the *sameAs.cc* data set [1] as our identity network $\mathcal{G}_{\sim}$. This data set contains all 556M non-reflexive `owl:sameAs` statements available in the *LOD-a-lot*, in addition to their resulting non-singleton 48.9M equivalence classes after transitive closure. The largest equivalence class contains 177K nodes, whilst 64% of these classes are of size 2. The

[2] http://lod-a-lot.lod.labs.vu.nl
[3] http://krr.triply.cc/krr/lod-a-lot

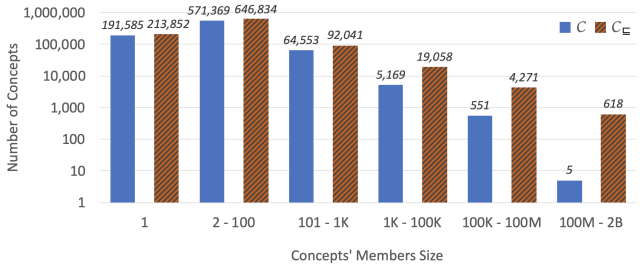

Figure 1: Size distribution of the concepts' members in the *LOD-a-lot* data set. Blue bins refer to the size of the concepts' explicit members, whilst brown/striped bins refer to the size of the concepts' both explicit and implicit members.

sameAs.cc data set is exposed in a single HDT file that is 5GB in size, and is publicly accessible via an LDF interface and a SPARQL client through the *sameAs.cc* identity web service[4]. The equivalence classes are exposed in two CSV files, which we convert into two RocksDB key-value stores using the RocksDB Python API[5]. These two key-value stores have the following structure:

- $[i]^{ID} \mapsto [i]$: in this file each equivalence class $[i]$, composed of a set of identical nodes, is associated with a unique identifier $[i]^{ID}$.
- $v \mapsto [v]^{ID}$: in this file each node v in $\mathcal{G}_{\sim}$ is mapped to its corresponding equivalence class identifier[6].

Concepts. The *LOD-a-lot* data set contains over 3.3B `rdf:type` statements. There is over 833K distinct concepts that appear in the object position of an `rdf:type` statement (i.e., $|C|$). There is an additional 143K concepts which members can only be deduced after exploiting the transitive closure of the subsumption relation (via the `rdfs:subClassOf` relation) which we denote by $|C_{\sqsubseteq}|$. Figure 1 presents the size distribution of these concepts' explicit and implicit members. It shows that most concepts have relatively few instances as members, with around 23% of the concepts appearing as objects in solely one `rdf:type` statement, and around 92% appearing as objects in less than 100 `rdf:type` statements. This figure also shows that the number of concepts with more than 100M members significantly increases when members are also deduced via the closure of the `rdfs:subClassOf` relation (increases from 5 to 618 concepts). Table 2 shows the only five concepts having more than 100M explicit members. This Table also shows that around 62% of the `rdf:type` statements in the *LOD-a-lot* data set have one of these five concepts in the object position.

5 EVALUATION

In this section, we use the *LOD-a-lot* and *sameAs.cc* data sets to provide empirical answers on our research questions: (Q1) whether considering `owl:sameAs` links can help improving the quality of schema alignments (*w* and *w/o* considering the transitive closure of the class subsumption relation); (Q2) whether there is a correlation between the `owl:sameAs` links' quality and the resulting alignments. To this end, we rely on existing alignments in the *LOD-a-lot* data set for constructing our benchmark, by making the assumption that all asserted alignments are correct. Available benchmarks constructed as part of the OAEI[7] campaigns (Ontology Alignment Evaluation Initiative) cannot be deployed for this evaluation, as they either take part of synthetically generated data sets or they were not covered in the LOD Laundromat 2015 crawl.

From the *LOD-a-lot* data set, we extract all 1,051,979 concept alignments (i.e., `owl:equivalenceClass` statements). Since our study relies on the presence of the concepts' extension for measuring the impact of `owl:sameAs` on schema alignments, we discard all alignments in which at least one of the aligned concepts have no explicit members. Out of the remaining 972 alignments, we also discard the 208 reflexive alignments and the 22 duplicate symmetric alignments. This results in a benchmark of 742 alignments, between 1,357 distinct concepts. The concept with the highest number of explicit members in this benchmark is `foaf:Person` i.e., $|ext(\texttt{foaf:Person})| \simeq 132M$, and the three concepts having almost equally the highest number of members after subsumption are `rdfs:Class`, `owl:Class`, and `owl:Thing` with more than 469M implicit members each. Figure 2 shows the size distribution of the concepts' members in this benchmark. From this Figure, we can observe that the concepts included in this benchmark have a similar size distribution to the full set of concepts in the *LOD-a-lot data set*, presented in Figure 1.

The evaluation conducted for investigating whether `owl:sameAs` enhances instance-based schema alignments (first research question) is twofold: firstly we investigate in Section 5.1 whether `owl:sameAs` increases the overlap of equivalent concepts, which are the 742 alignments in our benchmark; and secondly we investigate in Section 5.2, whether `owl:sameAs` have similar impact on non-equivalent concepts. The second research question is addressed in Section 5.3. All the raw results and the necessary data and scripts for replicating these evaluations are available at https://github.com/raadjoe/impact-sameAs-schema-matching.

Table 2: The only five concepts that appear in the object position of more than 100M `rdf:type` statements in the *LOD-a-lot*.

Concept	Cardinality	%
http://purl.org/linked-data/cube#Observation	1,306,389,396	39.3
http://data-gov.tw.rpi.edu/2009/data-gov-twc.rdf#DataEntry	304,878,654	9.2
http://geovocab.org/geometry#Geometry	167,808,111	5
http://knoesis.wright.edu/ssw/ont/sensor-observation.owl#MeasureData	144,044,989	4.3
http://xmlns.com/foaf/0.1/Person	132,919,327	4
Total	**2,056,040,477**	**61.9**

[4] http://sameas.cc
[5] See http://rocksdb.org/ and http://github.com/twmht/python-rocksdb
[6] we note that since equivalence classes form a partitioning of the nodes in $\mathcal{G}_{\sim}$, each node belongs to one unique equivalence class.
[7] http://oaei.ontologymatching.org

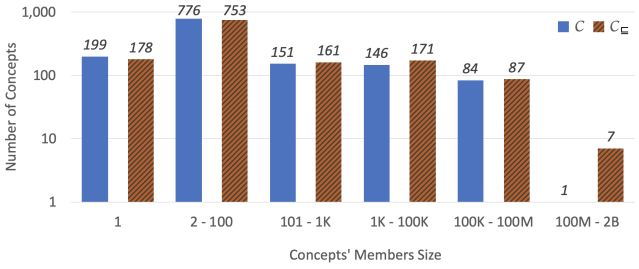

Figure 2: Size distribution of the concepts' members of our benchmark in the *LOD-a-lot* data set. Blue bins refer to the size of the concepts' explicit members, whilst brown/striped bins refer to the size of the concepts' both explicit and implicit members.

5.1 Does owl:sameAs increase the Jaccard index of equivalent concepts?

Evaluating whether the inclusion of owl:sameAs links increases the Jaccard index is straightforward. Similarly to Example 4.1, we measure J of the aligned concepts' instance sets without owl:sameAs links, and compare it to J after including owl:sameAs links. In addition, in order to investigate whether owl:sameAs has a different impact *with* or *without* the transitive closure of the class subsumption relation, we consider in the first part the concepts' explicit members only, before including their implicit members in the second part of the experiments.

Explicit Concept Members. For each pair of aligned concepts (C_1, C_2), we measure both their $J(ext(C_1), ext(C_2))$ and measure their $J(ext^{\sim}(C_1), ext^{\sim}(C_2))$, and observe how this Jaccard index varies. This process consists of (i) extracting the concepts' instance set, (ii) replacing each instance with its equivalence class identifier from the created RocksDB key-value store, and finally (iii) measuring their Jaccard index. The runtime of this process on the 742 alignments is ~90 minutes on an SSD disk, with 64GB of memory. Figure 3 presents the J distribution for these 742 alignments in our benchmark. It shows that indeed the inclusion of owl:sameAs links increases the J of equivalent concepts. In particular, we can observe that 322 pairs, previously with a Jaccard index of 0, have now a positive Jaccard index after including owl:sameAs links. In addition, we can observe that the number of pairs with a $J > 0.9$ has almost doubled when owl:sameAs was included. The mean Jaccard index of these 742 pairs increased from 0.07 to 0.222 when owl:sameAs links are considered.

Explicit & Implicit Concept Members. Similarly to the previous evaluation, we measure for each pair of aligned concepts (C_1, C_2) both their $J(ext_{\sqsubseteq}(C_1), ext_{\sqsubseteq}(C_2))$ and $J(ext^{\sim}_{\sqsubseteq}(C_1), ext^{\sim}_{\sqsubseteq}(C_2))$ for checking the impact of including owl:sameAs links also on implicit members. This process takes longer to finish (~ 4 hours), due to the increase in the number of concepts with large instance sets. Figure 4 presents the J distribution for the 742 aligned concepts of our benchmark when also implicit concept members are considered. The figure shows a slight increase of J when implicit members are also considered, both before and after considering owl:sameAs links. Finally, the mean Jaccard index of these 742 pairs increases from 0.08 to 0.223 when owl:sameAs links are considered.

Despite the average increase of the Jaccard index when owl:sameAs links are included, there is a total of 27 cases where considering owl:sameAs results in the decrease of the Jaccard index of two aligned concepts. Out of these 27 cases, there exists 23 cases that occur both when the concepts' only explicit members are considered, and when also their implicit members are considered, whereas two cases appear solely in the former, and two other cases appear only in the latter. Therefore, resulting in 25 cases each where J decreases, as Table 3 shows. Most of these cases occur in alignments between concepts from DBpedia and Schema.org, amounting in 19 out of 25 these cases (76%) when only explicit members are considered, and 17 cases (68%) when their implicit members are also considered. The largest decrease of J occurs between the concepts drugbank:Offer[8] and dailymed:Offer[9], where J decreases by 47% (from 0.46 to 0.24). Other than this case, the decrease of J is generally small: when only explicit members are considered, the average decrease is 0.026, with a median of 0.01; whereas the average decrease of J is 0.032, also with a median of 0.01 when both explicit and implicit members are considered.

From Table 3, we can also observe that when only explicit members of the concepts are considered, J increases for 361 pairs (49% of the cases) when owl:sameAs links are included. On the other hand, when both explicit and implicit members are considered, J increases for 381 pairs (52% of the cases). Thus, showing that in most cases, the inclusion of owl:sameAs links affects positively the Jaccard index of equivalent concepts, with a higher positive impact when also implicit members are considered. The mean increase when only explicit members are considered is 0.31, with a median of 0.19, whilst the mean when also implicit members are considered is 0.28, with a median of 0.13. This is mainly due to the 20 additional pairs that have a relatively small increase in their J, which affected both the mean and the median. Finally, Table 3 also shows that in 44 occasions (7% of the cases), the inclusion of owl:sameAs links increases the J of two equivalent concepts from 0 to 1. Interestingly, 42 out of these 44 cases (95%) are alignments between concepts from http://sw.opencyc.org/ and http://umbel.org/ namespaces.

5.2 Does owl:sameAs increase the Jaccard index of non-equivalent concepts?

In the previous section, we showed that when owl:sameAs links are considered, the Jaccard index of equivalent concepts in the *LOD-a-lot* data set increases in around half of the cases (between 49% and 52% depending if also implicit members are considered), and only decreases in 3% of the cases. In order to investigate whether owl:sameAs is indeed a positive factor for instance-based schema alignments techniques, we need to show that the inclusion of owl:sameAs does not increase the J of non-equivalent concepts. For this, we randomly pair all existing 833K concepts having at least one explicit member with each other, in a way that each concept is paired exactly once with another random concept. This results in ~416K new alignments, in which we assume that they

[8] http://www4.wiwiss.fu-berlin.de/drugbank/vocab/resource/class/Offer
[9] http://www4.wiwiss.fu-berlin.de/dailymed/vocab/resource/class/Offer

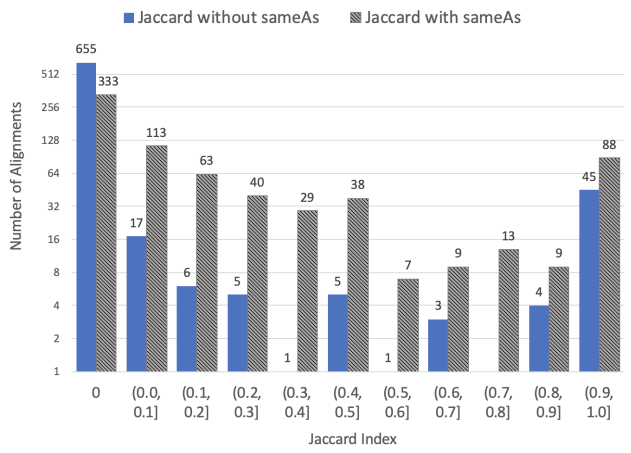

Figure 3: Jaccard Index distribution for the 742 alignments when only the concepts' explicit members are considered.

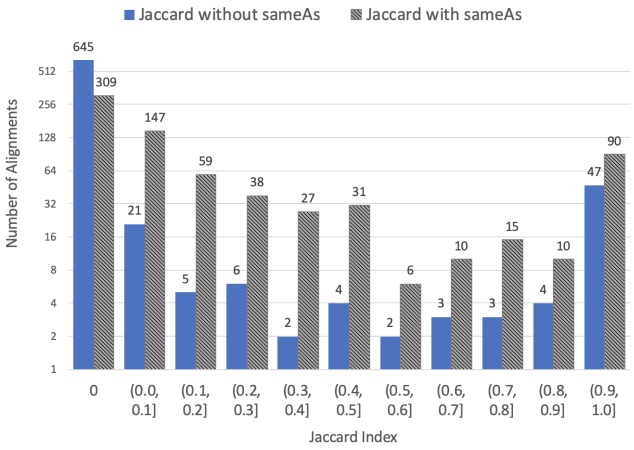

Figure 4: Jaccard Index distribution for the 742 alignments when both the concepts' explicit and implicit members are considered.

are all incorrect. Similarly to the previous evaluation, we measure for each pair of newly aligned pair of concepts (C_1, C_2) both their $J(ext_{\sqsubseteq}(C_1), ext_{\sqsubseteq}(C_2))$ and $J(ext^{\sim}_{\sqsubseteq}(C_1), ext^{\sim}_{\sqsubseteq}(C_2))$ for evaluating the impact of including owl:sameAs links on (most probably) incorrect alignments. The results of this experiment presented in Table 4, shows that out of these 416K randomly generated alignments, the inclusion of owl:sameAs links increases J for only 94 pairs of concepts (0.02% of the cases). This Table also shows that in 77 out of these 94 cases, the inclusion of owl:sameAs links have increased the J of different concepts from 0 to a positive value. However such increase of J is relatively small: average increase for these 94 cases is 0.008, with a median of 0.001, and a maximum increase of 0.14. The mean Jaccard index for these randomly aligned 416K pairs is 0.0033 and was not affected by the inclusion of owl:sameAs links due to its small increase in only 94 cases.

Table 3: Variation of J for the 742 aligned concepts when owl:sameAs is considered. The row 'Total' refers to the number of aligned pair of concepts with the corresponding J, prior to the consideration of owl:sameAs links.

	Jaccard Index	0	(0, 1)	1	Total
C	***Total***	***655 (88%)***	***73 (10%)***	***14 (2%)***	***742***
	Decreases	N/A	25 (34%)	0 (0%)	***25 (3%)***
	No variation	333 (51%)	9 (12%)	14 (100%)	***356 (48%)***
	Increases (J <1)	278 (42%)	39 (54%)	N/A	***317 (43%)***
	Increases (J = 1)	44 (7%)	0 (0%)	N/A	***44 (6%)***
$C_{\sqsubseteq}$	***Total***	***645 (87%)***	***81 (11%)***	***16 (2%)***	***742***
	Decreases	N/A	25 (31%)	0 (0%)	***25 (3%)***
	No variation	309 (48%)	11 (14%)	16 (100%)	***336 (45%)***
	Increases (J <1)	292 (45%)	45 (55%)	N/A	***337 (46%)***
	Increases (J = 1)	44 (7%)	0 (0%)	N/A	***44 (6%)***

Table 4: Variation of J for the 416K randomly aligned concepts when owl:sameAs is considered. The row 'Total' refers to the number of aligned pair of concepts with the corresponding J, prior to the consideration of owl:sameAs links.

Jaccard Index	0	(0, 1)	1	Total
Total	***412,828 (99.1%)***	***2,808 (0.67%)***	***980 (0.23%)***	***416,616***
Decreases	N/A	3 (0.1%)	0 (0%)	***3 (0%)***
No variation	412,751 (99.98%)	2,788 (99.3%)	980 (100%)	***416,519 (99.98%)***
Increases (J <1)	77 (0.02%)	17 (0.6%)	N/A	***94 (0.02%)***
Increases (J = 1)	0 (0%)	0 (0%)	N/A	***0 (0%)***

5.3 Does the quality of owl:sameAs links impact the quality of the alignments?

In the two previous sections, we showed that considering all existing owl:sameAs links in the *LOD-a-lot* data set increases J for 52% of the existing alignments, and decreases J for 3% of the alignments when concepts' implicit members are also considered. In addition, by randomly generating 416K alignments between 833K concepts, we showed that considering all owl:sameAs links increases the Jaccard index of 94 randomly aligned pair of concepts (0.02% of the cases). Following a number of studies showing that owl:sameAs is misused in the Web of data [8, 10, 17], we investigate in this section whether selecting a subset of these owl:sameAs links, of higher quality, can enhance the results presented in the previous sections. Ideally, deploying a curated collection of owl:sameAs links for measuring the Jaccard index of a pair of concepts' members, we expect mainly to prevent the decrease of J for the 25 correct alignments, and prevent the increase of J for the 94 incorrect alignments.

For selecting a higher quality subset of owl:sameAs links, we rely on the recent approach by [17] conducted also on the *sameAs.cc* data set. In this work, the authors computed an error degree between 0 and 1 for each of the existing 558M owl:sameAs statements, relying solely on the community structure of the identity network and the symmetrical property of the links. It is based on the assumption that the more an owl:sameAs link is isolated in the identity network, the higher the probability that it might be erroneous. This work shows that only by discarding the 1M owl:sameAs with an error degree higher than 0.99, the correctness of the resulting equivalence classes significantly increases, while at the same time limiting the number of truly identical instances that are separated from the same equivalence class. Furthermore, this study also shows that by considering only the 400M owl:sameAs links with an error degree lower or equal to 0.4, the newly resulted equivalence classes become almost 100% correct, based on the manual evaluation of 15K links. However in this case, when over 150M owl:sameAs links with an error degree higher than 0.4 results are discarded, a number of truly identical instances are separated into different (in most cases singleton) equivalence classes.

In this section, we use these results for conducting two separate experiments for measuring the impact of the owl:sameAs links' quality on the schema alignments. The first experiment (a) considers the equivalence classes resulted from the closure of the 557M owl:sameAs with an error degree <0.99, whilst the second experiment (b) considers the equivalence classes resulted from the closure of the 400M owl:sameAs with an error degree ≤0.4. Similarly to the process conducted in Section 4.2 on the original equivalence classes, these resulted equivalence classes from both closures, are converted from CSV files into separate RocksDB key-value stores for efficient access.

Impact of owl:sameAs quality on correct alignments. In this first part of the experiment, we investigate the impact of considering these higher quality subsets of owl:sameAs on the Jaccard index of the 742 pairs in our benchmark. Thus, for each pair of aligned concepts (C_1, C_2), we measure both their $J(ext_{\sqsubseteq}(C_1), ext_{\sqsubseteq}(C_2))$ and $J(ext^{\sim}_{\sqsubseteq}(C_1), ext^{\sim}_{\sqsubseteq}(C_2))$ by (a) considering only owl:sameAs links with error degree <0.99, and (b) considering only links with error degree ≤0.4. The results of these two separate experiments are presented in Table 5. These results shows worse results compared to the results previously presented in Table 3 when all owl:sameAs links were considered. Firstly, when owl:sameAs links with an error degree ≥0.99 are discarded, the number of pairs in the benchmark having an increase of J from 0 to 1 drops from 44 (7%) to 37 (6%), and the total number of pairs having their J increased in general slightly drops from 381 (52% of all pairs) to 376 (51%). The mean Jaccard index of all 742 pairs in our benchmark slightly decreases from 0.223 to 0.22. On the other hand, when owl:sameAs links with an error degree ≥0.4 are discarded, the positive impact of owl:sameAs on the J of the equivalent pairs of our benchmark is significantly reduced. Specifically, the number of equivalent concepts in the benchmark having an increase of J from 0 to 1 drops from 44 (7% of the pairs) to 2 (0.3%). In addition, the total number of pairs having their J increased in general drops from 381 (52% of the pairs) to 98 (12.9%), and the mean Jaccard index of all 742 pairs in our benchmark decreases in this case from 0.223 to 0.094.

Table 5: Variation of J for the 742 aligned concepts when (a) only owl:sameAs links with error degree < 0.9 are considered and (b) when only owl:sameAs links with error degree < 0.4 are considered. The row 'Total' refers to the number of aligned pair of concepts with the corresponding J, prior to the consideration of owl:sameAs links.

	Jaccard Index	0	(0, 1)	1	Total
$C_{\sqsubseteq}$ (a)	***Total***	***645 (87%)***	***81 (11%)***	***16 (2%)***	***742***
	Decreases	N/A	25 (31%)	0 (100%)	***25 (3%)***
	No variation	313 (48%)	12 (15%)	16 (100%)	***341 (46%)***
	Increases (J <1)	295 (46%)	44 (54%)	N/A	***339 (46%)***
	Increases (J = 1)	37 (6%)	0 (0%)	N/A	***37 (5%)***
$C_{\sqsubseteq}$ (b)	*Decreases*	N/A	39 (48%)	0 (0%)	***39 (5.2%)***
	No variation	570 (88.4%)	19 (24%)	16 (100%)	***605 (81.5%)***
	Increases (J <1)	73 (11.3%)	23 (28%)	N/A	***96 (12.9%)***
	Increases (J = 1)	2 (0.3%)	0 (0%)	N/A	***2 (0.2%)***

Finally, one of the goals of this experiment is to test whether selecting a higher quality subset of owl:sameAs links would affect the 25 pairs of equivalent concepts having their J decreased. The results from Table 5 shows that these 25 cases remain in both experiments (a) and (b). On the opposite, an additional 14 cases occurs in experiment (b), where the J of equivalent pairs of concepts have decreased. However, the average decrease of J for these 25 pairs of aligned concepts drops from 0.032 to 0.028 in experiment (a), and drops to 0.012 in experiment (b).

Impact of owl:sameAs quality on random alignments. The previously presented experiments on the 742 equivalent pairs in our benchmark have shown a slight negative decrease of impact when only owl:sameAs links with error degree <0.99 are considered compared to considering all owl:sameAs links, and a significant negative decrease of impact when only links with error degree ≤0.4 are considered. In this section, we investigate whether considering these same subsets of owl:sameAs links have a different impact on the 416K random alignments generated in Section 5.2. Ideally, we expect by considering a higher quality subsets of owl:sameAs links, to reduce the number of randomly aligned pairs with an increased J. The results presented in Table 6 indeed shows that the higher the quality of the considered collection of owl:sameAs is, the less frequent an increase of J occurs between non-equivalent pair of concepts. Specifically, when only owl:sameAs links with an error degree <0.99 are considered, the number of incorrect alignments with an increase in J drops from 94 to 27 (71% improvement). Whereas, when only owl:sameAs links with an error degree ≤0.4 are considered, the number of incorrect alignments with an increase in J drops from 94 to 2 (98% improvement).

Table 6: Variation of J for the 416K randomly aligned concepts when (a) only `owl:sameAs` links with error degree < 0.9 are considered and (b) when only `owl:sameAs` links with error degree < 0.4 are considered. The row 'Total' refers to the number of aligned pair of concepts with the corresponding J, prior to the consideration of `owl:sameAs` links.

	Jaccard Index	0	(0, 1)	1	Total
(a)	***Total***	***412,828 (99.1%)***	***2,808 (0.67%)***	***980 (0.23%)***	416,616
	Decreases	N/A	3 (0.1%)	0 (0%)	***3 (~0%)***
	No variation	412,817 (~100%)	2,789 (99.3%)	980 (100%)	***416,586 (~100%)***
	Increases (J <1)	11 (~0%)	16 (0.6%)	N/A	***27 (~0%)***
	Increases (J = 1)	0 (0%)	0 (%)	N/A	***0 (0%)***
(b)	*Decreases*	N/A	0 (0%)	0 (0%)	***0 (0%)***
	No variation	412,828 (100%)	2,806 (99.93%)	980 (100%)	***416,614 (~100%)***
	Increases (J <1)	0 (0%)	2 (0.07%)	N/A	***2 (~0%)***
	Increases (J = 1)	0 (0%)	0 (0%)	N/A	***0 (0%)***

6 CONCLUSION

This paper presented an empirical study on the impact of considering `owl:sameAs` links in instance-based schema matching. This is the first study of this type and at this scale, enabled by the recent emergence of two important elements of infrastructure: the *LOD-a-lot* data set containing over 3 billion `rdf:type` statements, and the *sameAs.cc* data set containing over 35 billion identity links after closure. The main findings of this study are summarised as follows:

Including instance-level interlinks enhances instance-based schema alignments. Based on a benchmark of 742 equivalent pair of concepts extracted from the *LOD-a-lot* data set, the experiments conducted in Section 5.1 shows that the inclusion of `owl:sameAs` links increase the Jaccard index of around half of these pairs, with a decrease of Jaccard restricted to only 3% of these pairs. In addition, and based on a benchmark of 416K randomly generated alignments, the experiments conducted in Section 5.2 shows that including `owl:sameAs` links does not increase the Jaccard index of non-equivalent pairs, with an exception of 94 cases (0.02% of cases).

Inference does positively impact instance-based schema alignments. In addition of exploiting the transitive closure of `owl:sameAs`, exploiting the transitive closure of the subsumption relations in the Web also positively impacts instance-based schema matching. Specifically, the experiments conducted in Section 5.1, shows that considering also the concepts' implicit members increases the number of equivalent pair of concepts in our benchmark that have an increase in their Jaccard index, from 49% to 52%.

Discarding isolated `owl:sameAs` links can increase the quality of instance-based schema alignments. The experiments conducted in Section 5.3 shows that discarding ~1M `owl:sameAs` that are isolated in the network (links with error degree >0.99) reduces the probability of increasing the similarity of two non-equivalent concepts by 71%, without having an negative impact on the equivalent concepts in our benchmark.

We believe that the findings of this study can be of importance to the large ontology-matching community, as it provides empirical evidences on the benefits of using external collection of instance-level interlinks for their task of linking multiple schemas. Building on the findings of this study, we will further investigate other better-tailored instance-based measures, which can exploit the curated collection of `owl:sameAs` links and the implicit members of the concepts, in order to detect new alignments at the scale of the Web. This will require making different technical choices for reducing the runtime of the process, which is mainly affected by the search in the key-value store for each member, when comparing each pair of concepts.

REFERENCES

[1] Wouter Beek, Joe Raad, Jan Wielemaker, and Frank van Harmelen. 2018. sameAs.cc: The Closure of 500M owl: sameAs Statements. In *Extended Semantic Web Conference*. Springer, 65–80.

[2] Wouter Beek, Laurens Rietveld, Hamid R Bazoobandi, Jan Wielemaker, and Stefan Schlobach. 2014. LOD laundromat: a uniform way of publishing other people's dirty data. In *International Semantic Web Conference*. Springer, 213–228.

[3] Gianluca Correndo, Antonio Penta, Nicholas Gibbins, and Nigel Shadbolt. 2012. Statistical analysis of the owl: sameAs network for aligning concepts in the linking open data cloud. In *International Conference on Database and Expert Systems Applications*. Springer, 215–230.

[4] AnHai Doan, Jayant Madhavan, Pedro Domingos, and Alon Halevy. 2004. Ontology matching: A machine learning approach. In *Handbook on ontologies*. Springer, 385–403.

[5] Jérôme Euzenat and Pavel Shvaiko. 2013. Ontology Matching, Second Edition.

[6] Javier D Fernández, Wouter Beek, Miguel A Martínez-Prieto, and Mario Arias. 2017. LOD-a-lot. In *International Semantic Web Conference*. Springer, 75–83.

[7] Hugh Glaser, Afraz Jaffri, and Ian Millard. 2009. Managing Co-reference on the Semantic Web. In *Proceedings of the WWW Workshop on Linked Data on the Web, LDOW*.

[8] Harry Halpin, Patrick J Hayes, James P McCusker, Deborah L McGuinness, and Henry S Thompson. 2010. When owl:sameAs isn't the same: An analysis of identity in Linked Data. In *International Semantic Web Conference*. Springer, 305–320.

[9] Pascal Hitzler, Markus Krotzsch, and Sebastian Rudolph. 2009. *Foundations of semantic web technologies*. Chapman and Hall/CRC.

[10] Aidan Hogan, Antoine Zimmermann, Jürgen Umbrich, Axel Polleres, and Stefan Decker. 2012. Scalable and distributed methods for entity matching, consolidation and disambiguation over linked data corpora. *Web Semantics: Science, Services and Agents on the World Wide Web* 10 (2012), 76–110.

[11] Antoine Isaac, Lourens Van Der Meij, Stefan Schlobach, and Shenghui Wang. 2007. An empirical study of instance-based ontology matching. In *The Semantic Web*. Springer, 253–266.

[12] Afraz Jaffri, Hugh Glaser, and Ian Millard. 2008. URI Disambiguation in the Context of Linked Data. In *WWW Workshop on Linked Data on the Web, LDOW*.

[13] Michael Levandowsky and David Winter. 1971. Distance between sets. *Nature* 234, 5323 (1971), 34.

[14] Michalis Mountantonakis and Yannis Tzitzikas. 2016. On measuring the lattice of commonalities among several linked datasets. *Proceedings of the VLDB Endowment* 9, 12 (2016), 1101–1112.

[15] Andriy Nikolov, Victoria Uren, Enrico Motta, and Anne De Roeck. 2009. Overcoming schema heterogeneity between linked semantic repositories to improve coreference resolution. In *Asian Semantic Web Conference*. Springer, 332–346.

[16] Joe Raad. 2018. Identity Management in Knowledge Graphs (doctoral dissertation). University of Paris-Saclay.

[17] Joe Raad, Wouter Beek, Frank van Harmelen, Nathalie Pernelle, and Fatiha Saïs. 2018. Detecting Erroneous Identity Links on the Web Using Network Metrics. In *International Semantic Web Conference*. Springer, 391–407.

[18] Gerd Stumme and Alexander Maedche. 2001. FCA-Merge: Bottom-up merging of ontologies. In *IJCAI*, Vol. 1. 225–230.

[19] Fabian M Suchanek, Serge Abiteboul, and Pierre Senellart. 2011. Paris: Probabilistic alignment of relations, instances, and schema. *Proceedings of the VLDB Endowment* 5, 3 (2011), 157–168.