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.archives-ouvertes.fr/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 \texttt{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 \texttt{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:

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 \texttt{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. \texttt{rdfs:seeAlso}, \texttt{owl:differentFrom}), this work considers only equivalence relations found in the form of \texttt{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 \texttt{owl:sameAs} links are actually erroneous [8, 10, 17]. This uncertainty regarding the quality of existing \texttt{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 \texttt{sameAs.cc} identity-cloud...
[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 sets 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].

Another approach is combining both instance- and schema-level information for a single system. A notable example of this is the K-CAP’19, November 2019, Marina del Rey, California, USA

Joe Raad, Erman Acar, and Stefan Schlobach

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 sets 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].

Another approach is combining both instance- and schema-level information for a single system. A notable example of this is the K-CAP’19, November 2019, Marina del Rey, California, USA

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 sets 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].

Another approach is combining both instance- and schema-level information for a single system. A notable example of this is the K-CAP’19, November 2019, Marina del Rey, California, USA

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 sets 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].

Another approach is combining both instance- and schema-level information for a single system. A notable example of this is the K-CAP’19, November 2019, Marina del Rey, California, USA

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

The set of all concepts we consider are those that appear in the object positions of an RDF triple $(i, \text{rdfs:typeof}, C) \in \mathcal{K}$ with $i$ called an instance, and is denoted by calligraphic $C$. By $\mathcal{C}(i)$, we denote the set of concepts whose $i$ is a member. Similar to the aforementioned notions of extensions, $C_1^i(i)$, $C_2^i(i)$, and $C_5^i(i)$ 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 $\text{owl:sameAs}$ 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 $\text{owl:sameAs}$ 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 $\text{owl:sameAs}$ 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 $\text{ext}(C_1) = \{ i_1, i_2, i_3, i_4 \}$, and $\text{ext}(C_2) = \{ i_1, i_2, i_5 \}$. With $\text{ext}(C_1) \cap \text{ext}(C_2) = \{ i_1, i_2 \}$ and $\text{ext}(C_1) \cup \text{ext}(C_2) = \{ i_1, i_2, i_3, i_4, i_5 \}$, the resulting $J(\text{ext}(C_1), \text{ext}(C_2))$ 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 $i_2$ and $i_5$ from the previous example, i.e., $\langle i_2, \text{owl:sameAs}, i_5 \rangle$, hence both $i_2$ and $i_5$ belong to the same equivalence class $[i]$. In this scenario, replacing all instances that belong to the same equivalence class with a unique identifier $[i]$ results in $\text{ext}^*(C_1) \cup \text{ext}^*(C_2) = \{ i_1, i_2, [i]^{ID}, i_5 \}$. With the decrease of their union size, while their intersection stays invariant, $J(\text{ext}^*(C_1), \text{ext}^*(C_2))$ 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., $\langle i_3, \text{owl:sameAs}, i_4 \rangle$. In such scenario, $|\text{ext}^*(C_1) \cap \text{ext}^*(C_2)|$ increases and $|\text{ext}^*(C_1) \cup \text{ext}^*(C_2)|$ decreases, resulting in a higher increase of $J(\text{ext}^*(C_1), \text{ext}^*(C_2))$ 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, $J(\text{ext}^*(C_1), \text{ext}^*(C_2))$ 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 $\text{owl:sameAs}$ 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.

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.

<table>
<thead>
<tr>
<th># triples</th>
<th>28,362,198,927</th>
</tr>
</thead>
<tbody>
<tr>
<td># rdf:type</td>
<td>3,321,354,308</td>
</tr>
<tr>
<td># owl:sameAs</td>
<td>558,943,116</td>
</tr>
<tr>
<td># equivalence classes</td>
<td>48,999,148</td>
</tr>
<tr>
<td># rdfs:subClassOf</td>
<td>4,461,717</td>
</tr>
<tr>
<td># owl:equivalentClass</td>
<td>1,051,979</td>
</tr>
<tr>
<td></td>
<td>[C</td>
</tr>
<tr>
<td></td>
<td>976,674</td>
</tr>
</tbody>
</table>

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

**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 HDI file2 that is 524GB in size, and is publicly accessible via an LDF (Linked Data Fragments) interface3.

**Identity Network & Equivalence Classes.** We use the sameAs.cc data set [1] as our identity network $G_\text{--}$. This data set contains all 556M non-reflexive $\text{owl:sameAs}$ 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

2http://lod-a-lot.lod.labs.vu.nl
3http://krr.triple.cc/krr/lod-a-lot
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. The equivalence classes are exposed in two CSV files, which we convert into two RocksDB key-value stores using the RocksDB Python API. These two key-value stores have the following structure:

- $[i]^{ID}$ in this file each equivalence class $[i]$, composed of a set of identical nodes, is associated with a unique identifier $[i]^{ID}$.
- $v \rightarrow [v]^{ID}$ in this file each node $v$ in $G_{C}$ is mapped to its corresponding equivalence class identifier.

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 rdf:s:subClassOf relation) which we denote by $|C|_e$. 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 rdf:s:subClassOf relation (increases from 5 to 618 concepts). Table 2 shows the only five concepts having more than 100M explicit members, whilst brown/striped bins refer to the size of the concepts’ both explicit and implicit members.

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.

<table>
<thead>
<tr>
<th>Concept</th>
<th>Cardinality</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://geovocab.org/geometry#Geometry">http://geovocab.org/geometry#Geometry</a></td>
<td>167,808,111</td>
<td>5</td>
</tr>
<tr>
<td><a href="http://knoesis.wright.edu/sw/ont/sensor-observation.owl#MeasureData">http://knoesis.wright.edu/sw/ont/sensor-observation.owl#MeasureData</a></td>
<td>144,044,989</td>
<td>4.3</td>
</tr>
<tr>
<td><a href="http://xmlns.com/foaf/0.1/Pet">http://xmlns.com/foaf/0.1/Pet</a></td>
<td>132,919,327</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>2,056,040,477</td>
<td>61.9</td>
</tr>
</tbody>
</table>

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.
previous evaluation, we measure for each pair of aligned concepts
varies. This process consists of (i) extracting the concepts’ instance
members are also considered, both before and after considering
$J$ due to the increase in the number of concepts with large instance
	\[ \sim p \]
addition, we can observe that the number of pairs with a
$J$ now a positive Jaccard index after including
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 $J(\text{ext}(C_1), \text{ext}(C_2))$ and measure their
$J(\text{ext}^-(C_1), \text{ext}^-(C_2))$, 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 $\text{owl:sameAs}$ 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 $\text{owl:sameAs}$ links. In addition, we can observe that the number of pairs with a $J > 0.9$ has almost doubled when $\text{owl:sameAs}$ was included. The mean Jaccard index of these 742 pairs increased from 0.07 to 0.222 when $\text{owl:sameAs}$ 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 $J(\text{ext}(C_1), \text{ext}(C_2))$ and $J(\text{ext}^-(C_1), \text{ext}^-(C_2))$ for checking the impact of including $\text{owl:sameAs}$ 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

\[ \text{owl:sameAs} \]

Despite the average increase of the Jaccard index when $\text{owl:sameAs}$ links are included, there is a total of 27 cases where considering $\text{owl:sameAs}$ 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 and da1ymed:Offer, 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 $\text{owl:sameAs}$ 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 $\text{owl:sameAs}$ 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 $\text{owl:sameAs}$ 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 $\text{owl:sameAs}$ increase the Jaccard index of non-equivalent concepts?**

In the previous section, we showed that when $\text{owl:sameAs}$ 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 $\text{owl:sameAs}$ is indeed a positive factor for instance-based schema alignments techniques, we need to show that the inclusion of $\text{owl:sameAs}$ 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

\[ http://www4.wiwiss.fu-berlin.de/drugbank/vocab/resource/class/Offer \]
\[ http://www4.wiwiss.fu-berlin.de/dailymed/vocab/resource/class/Offer \]
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 \(J(\text{ext}_C(C_1), \text{ext}_C(C_2))\) and \(J(\text{ext}_C^\prime(C_1), \text{ext}_C^\prime(C_2))\) for evaluating the impact of including \(\text{owl:sameAs}\) 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 \(\text{owl:sameAs}\) 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 \(\text{owl:sameAs}\) 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, the inclusion of \(\text{owl:sameAs}\) links increases \(J\) 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 \(\text{owl:sameAs}\) links increases the Jaccard index of 94 randomly aligned pair of concepts (0.02% of the cases). Following a number of studies showing that \(\text{owl:sameAs}\) is misused in the Web of data [8, 10, 17], we investigate in this section whether selecting a subset of these \(\text{owl:sameAs}\) links, of higher quality, can enhance the results presented in the previous sections. Ideally, deploying a curated collection of \(\text{owl:sameAs}\) 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.

<table>
<thead>
<tr>
<th>Jaccard Index</th>
<th>0</th>
<th>(0, 1)</th>
<th>1</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>(C)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>645</td>
<td>(87%)</td>
<td>91</td>
<td>16</td>
</tr>
<tr>
<td>Decreases</td>
<td>N/A</td>
<td>25</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>No variation</td>
<td>309</td>
<td>(48%)</td>
<td>11</td>
<td>16</td>
</tr>
<tr>
<td>Increases (J &lt; 1)</td>
<td>292</td>
<td>(45%)</td>
<td>45</td>
<td>N/A</td>
</tr>
<tr>
<td>Increases (J = 1)</td>
<td>44</td>
<td>(7%)</td>
<td>0</td>
<td>N/A</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Jaccard Index</th>
<th>0</th>
<th>(0, 1)</th>
<th>1</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>(C_e)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>412,828</td>
<td>(99.1%)</td>
<td>2,808</td>
<td>(0.67%)</td>
</tr>
<tr>
<td>Decreases</td>
<td>N/A</td>
<td>0</td>
<td>0</td>
<td>N/A</td>
</tr>
<tr>
<td>No variation</td>
<td>412,951</td>
<td>(99.98%)</td>
<td>2,788</td>
<td>(99.3%)</td>
</tr>
<tr>
<td>Increases (J &lt; 1)</td>
<td>17</td>
<td>(0.02%)</td>
<td>N/A</td>
<td>94</td>
</tr>
<tr>
<td>Increases (J = 1)</td>
<td>0</td>
<td>(0%)</td>
<td>N/A</td>
<td>0</td>
</tr>
</tbody>
</table>

5.3 Does the quality of \(\text{owl:sameAs}\) links impact the quality of the alignments?

In the two previous sections, we showed that considering all existing \(\text{owl:sameAs}\) 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 \(\text{owl:sameAs}\) links increases the Jaccard index of 94 randomly aligned pair of concepts (0.02% of the cases). Following a number of studies showing that \(\text{owl:sameAs}\) is misused in the Web of data [8, 10, 17], we investigate in this section whether selecting a subset of these \(\text{owl:sameAs}\) links, of higher quality, can enhance the results presented in the previous sections. Ideally, deploying a curated collection of \(\text{owl:sameAs}\) 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 (C1, C2), we measure both their \( j(\text{ext}_C(C_1), \text{ext}_C(C_2)) \) and \( j(\text{ext}^{-1}_C(C_1), \text{ext}^{-1}_C(C_2)) \) 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>
<thead>
<tr>
<th>Jaccard Index</th>
<th>0</th>
<th>(0, 1)</th>
<th>1</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Decreases</td>
<td>N/A</td>
<td>25 (31%)</td>
<td>0 (100%)</td>
<td>25 (3%)</td>
</tr>
<tr>
<td>No variation</td>
<td>313 (48%)</td>
<td>12 (15%)</td>
<td>16 (100%)</td>
<td>341 (446%)</td>
</tr>
<tr>
<td>Increases ( J = 1 )</td>
<td>295 (46%)</td>
<td>44 (54%)</td>
<td>N/A (446%)</td>
<td>339 (446%)</td>
</tr>
<tr>
<td>Increases ( J = 1 )</td>
<td>37 (6%)</td>
<td>0 (0%)</td>
<td>N/A (5%)</td>
<td>37 (5%)</td>
</tr>
</tbody>
</table>

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 pairs of concepts with the corresponding $J$, prior to the consideration of owl:sameAs links.

<table>
<thead>
<tr>
<th>Jaccard Index</th>
<th>0</th>
<th>(0, 1)</th>
<th>1</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total</strong></td>
<td>412,828</td>
<td>2,808</td>
<td>980</td>
<td>416,616</td>
</tr>
<tr>
<td><strong>Decreases</strong></td>
<td>N/A</td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td><strong>No variation</strong></td>
<td>412,828</td>
<td>2,808</td>
<td>980</td>
<td>416,616</td>
</tr>
<tr>
<td><strong>Increases</strong></td>
<td>(99.93%)</td>
<td>(99.3%)</td>
<td>(100%)</td>
<td>(99.1%)</td>
</tr>
<tr>
<td>$(J = 1)$</td>
<td>11</td>
<td>16</td>
<td>N/A</td>
<td>25</td>
</tr>
<tr>
<td>$(J &gt; 1)$</td>
<td>0</td>
<td>0</td>
<td>N/A</td>
<td>0</td>
</tr>
</tbody>
</table>

(b) | Decreases | N/A | 0 | 0 | 0 |
| No variation | 412,828 | 2,806 | 980 | 416,616 |
| $(J = 1)$ | 0 | 2 | N/A | 2 |
| $(J > 1)$ | 0 | 0 | N/A | 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 evidence 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