On the Impact of sameAs on Schema Matching

JOE RAAD ◉ ERMAN ACAR ◉ STEFAN SCHLOBACH

Journées RoD — July 6, 2020

(work published in KCAP 2019)

VU VRIJE UNIVERSITEIT AMSTERDAM

Knowledge Representation & Reasoning Group

More and more Linked Open Data...

krr.triply.cc/krr/lod-a-lot/

krr / lod-a-lot

Search ...

lod-a-lot

Browser

Table

Graphs 1

Services 0

Assets 0

Saved Queries 0

lod-a-lot

by Knowledge Representation & Reasoning (KRR)

Created 8 months ago
28.362.198.927 statements

(crawled from ~650K datasets in 2015)

LOD-a-lot is the graph merge of the RDF graphs that were part of the LOD Laundromat. LOD-a-lot was created by Fernández et al. 2017.

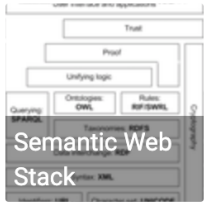
License

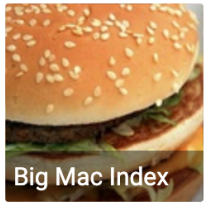
CC0 1.0

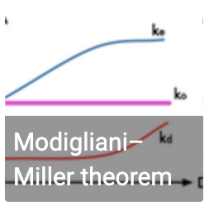
Graphs

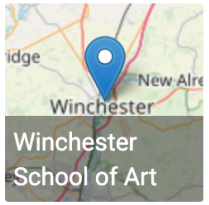
default 28.362.198.927

Example resources



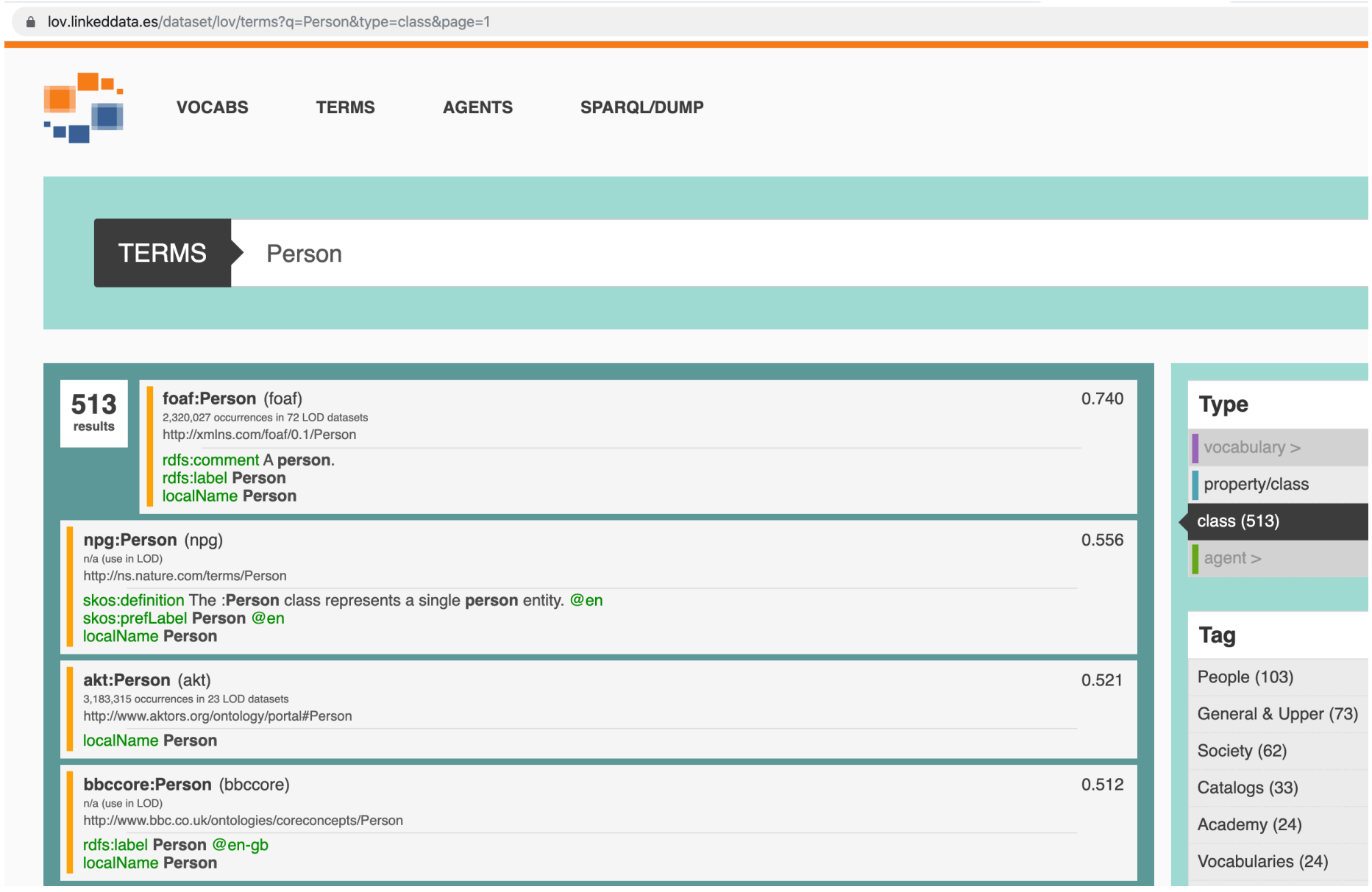






CommonKADS : 3rd KADS Meeting : Siemens AG Munich, March 8 - 9,1993 ; organized with GI special interest group 1.5.1 Knowledge Engineering

...More and more (overlapping) Schemas



Schema Matching is Inevitable

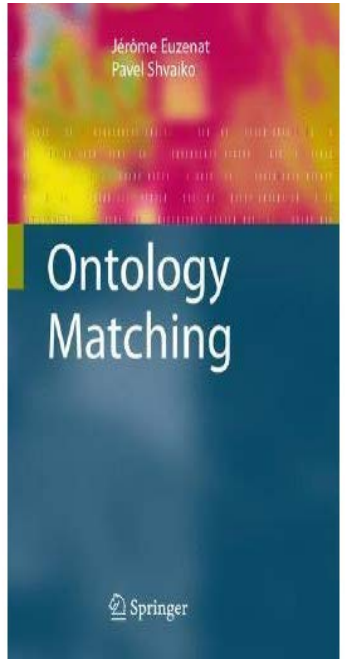
- It is not possible (neither desired) to have a unique schema covering all domains

- In order to exploit this wealth of available knowledge and enhance knowledge-based systems (e.g., search engines, virtual assistants, etc.), we need to match these overlapping schemas

- Schema Matching: finding relationships between entities of different schemas
  - equivalence relations
  - subsumption
  - disjointness
  - ....

Schema Matching over the years

- Active area of research from several communities, including the Semantic Web
- Ontology Alignment Evaluation Initiative (OAEI) ongoing for 15 years
- [Euzenat and Shvaiko, 2013] reviews ~100 schema-matching systems



50%
rely mostly on schema-level information
(i.e. schema-based approaches)

25%
rely on schema + instance-level information
(i.e. mixed approaches)

25%
rely mostly on instance-level information
(i.e. instance-based approaches)

Instance-based Schema Matching

- All instance-based schema-matching approaches share two essential ideas:

1. **The semantics of a concept is better determined by its members rather by its annotations**

Concepts refer to sets that possibly have named instances as ***members***

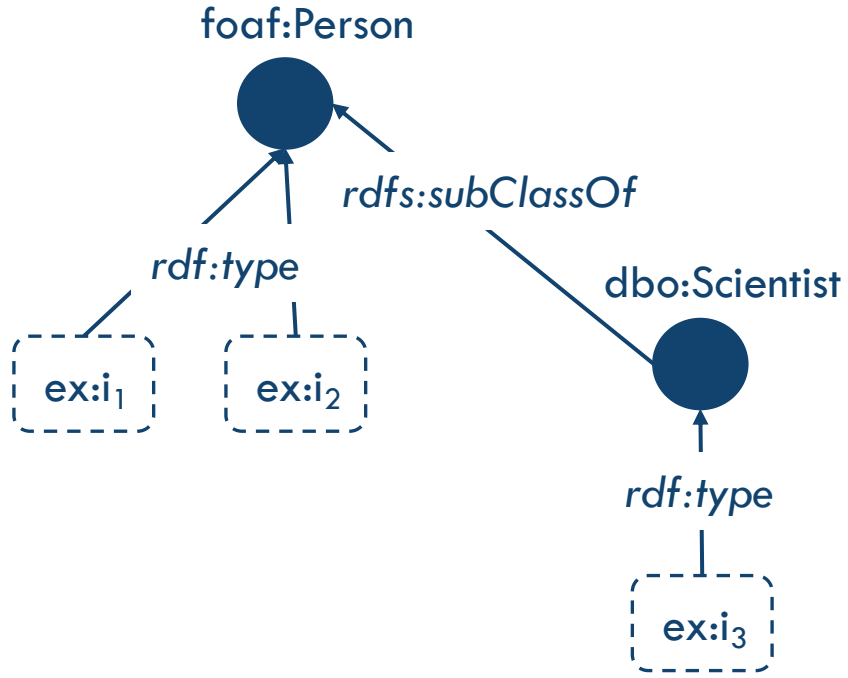
- ***ext(C)*** *refer to the set of instances which are* ***explicitly*** *stated as members of* ***C***

  $ext(foaf:Person) = \{ex:i_1, ex:i_2\}$

- ***ext$_{\sqsubseteq}$(C)*** *refer to the set of instances which are* ***explicitly*** *or* ***implicitly*** *stated as members of* ***C***

(i.e. either explicit members or derived through concept subsumption)

$ext_{\sqsubseteq}(foaf:Person) = \{ex:i_1, ex:i_2, ex:i_3\}$



Instance-based Schema Matching

- All instance-based schema-matching approaches share two essential ideas:

2. **The more significant the overlap between two concepts' members is, the more related these concepts are**

  - Multiple techniques to measure the overlap between concepts' members
    - Formal concept analysis techniques
    - Machine learning
    - Jaccard index
    - ....

Instance-based Schema Matching using Jaccard Index

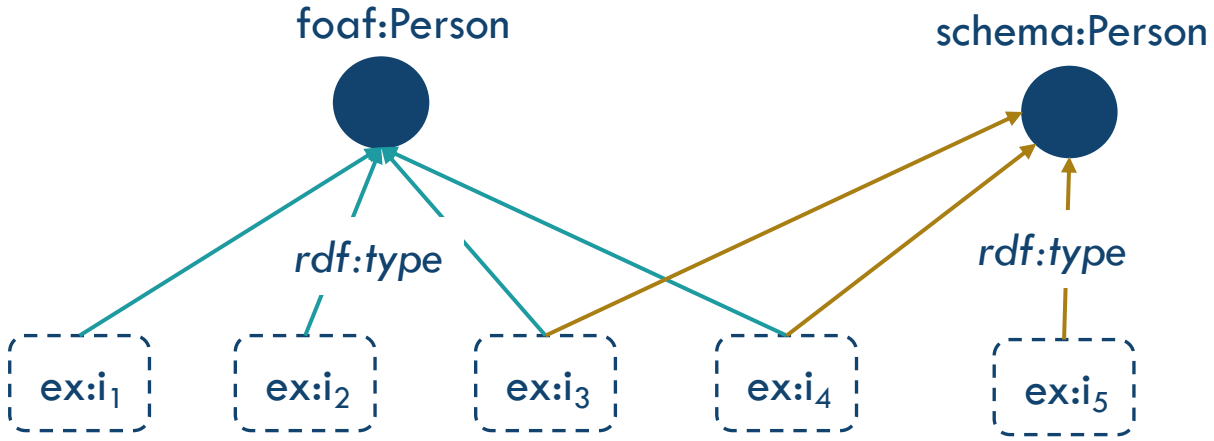
- The Jaccard index is a commonly used score to measure the similarity between two sets
- The higher the similarity of two sets is, the greater the Jaccard index

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

ext(foaf:Person) = {ex:i_1, ex:i_2 , ex:i_3 , ex:i_4}

ext(schema:Person) = {ex:i_3, ex:i_4, ex:i_5}

J(ext(foaf:Person), ext(schema:Person)) $= \frac{2}{5} = 0.4$



Instance-based Schema Matching using Jaccard Index

With more than 558 million explicitly asserted owl:sameAs [Beek et al., ESWC 2018]

(or 35 billion after transitive closure), the reality in the Web of Data looks more like this:

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

Scenario 1 where J increases

ext(foaf:Person) = {ex:i_1, ex:i_2 , ex:i_3 , ex:i_4}

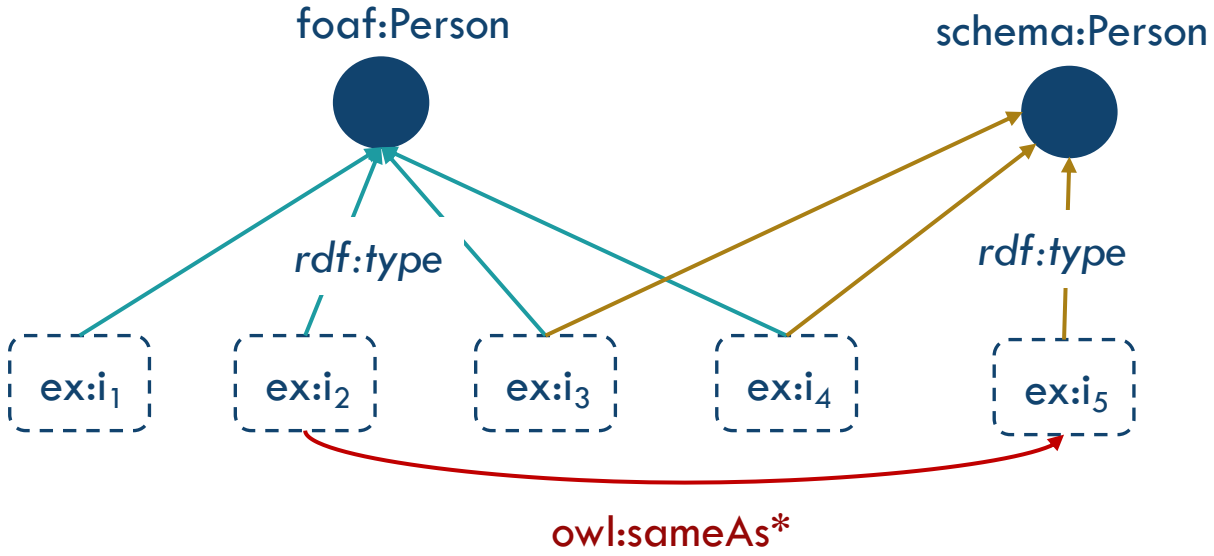
ext(schema:Person) = {ex:i_3, ex:i_4, ex:i_5}

J(ext(foaf:Person), ext(schema:Person)) $= \frac{2}{5} = 0.4$

ext~(foaf:Person) = {ex:i_1, $eq^{\{2,5\}}$, ex:i_3, ex:i_4}

ext~(schema:Person) = {ex:i_3, ex:i_4, $eq^{\{2,5\}}$}

J(ext~(foaf:Person), ext~(schema:Person)) $= \frac{3}{4} = 0.75$



Instance-based Schema Matching using Jaccard Index

Or possibly like this:

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

Scenario 2 where J decreases

ext(foaf:Person) = {ex:i_1, ex:i_2 , ex:i_3 , ex:i_4}

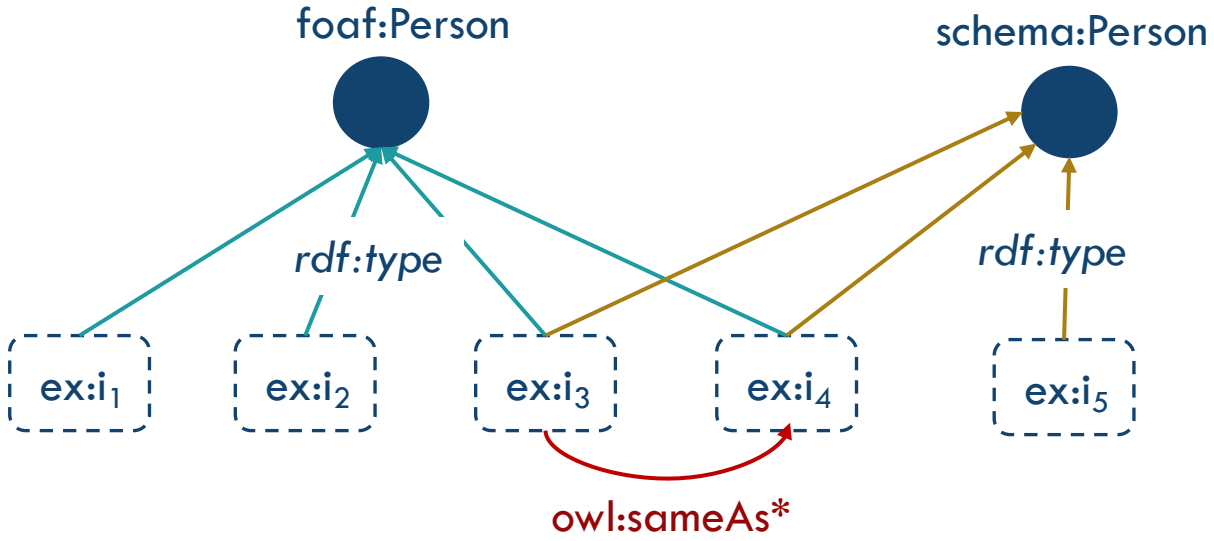
ext(schema:Person) = {ex:i_3, ex:i_4, ex:i_5}

J(ext(foaf:Person), ext(schema:Person)) $= \frac{2}{5} = 0.4$

ext~(foaf:Person) = {ex:i_1, ex:i_2, eq$^{\{3,4\}}$}

ext~(schema:Person) = {eq$^{\{3,4\}}$, ex:i_5}

J(ext~(foaf:Person), ext~(schema:Person)) $= \frac{1}{4} = 0.25$



Research Question

Research Question

Does the inclusion of instance-level interlinks (i.e. owl:sameAs) positively impact instance-based schema alignments ?

a. Does the inclusion of owl:sameAs increase the Jaccard Index of equivalent concepts?

b. Does the inclusion of owl:sameAs increase the Jaccard Index of non-equivalent concepts?

Why should we care?

- Provides empirical evidence for schema-matching designers on whether exploiting a large external collection of instance-level interlinks (e.g. from the LOD Cloud) is beneficial for improving the accuracy of schema-matching techniques

- Shows the risks/benefits of using owl:sameAs after a number of studies suggesting that a large* number of the existing owl:sameAs links in the Web are actually erroneous

  * **20%** of evaluated owl:sameAs are erroneous [Halpin et al., ISWC 2010]
  * **3%** of evaluated owl:sameAs are erroneous [Hogan et al., JWS 2012]
  * **4%** of evaluated owl:sameAs are erroneous [Raad et al., ISWC 2018]

Dataset Description

Dataset

krr.triply.cc/krr/lod-a-lot/

krr / lod-a-lot

Search ...

- lod-a-lot
- Browser
- Table
- Graphs 1
- Services 0
- Assets 0
- Saved Queries 0

lod-a-lot

by Knowledge Representation & Reasoning (KRR)

Created 8 months ago
28.362.198.927 statements

(crawled from ~650K datasets in 2015)

LOD-a-lot is the graph merge of the RDF graphs that were part of the LOD Laundromat. LOD-a-lot was created by Fernández et al. 2017.

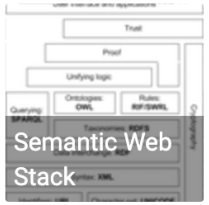
License

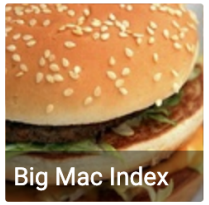
CC0 1.0

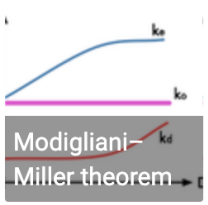
Graphs

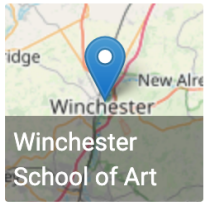
default 28.362.198.927

Example resources







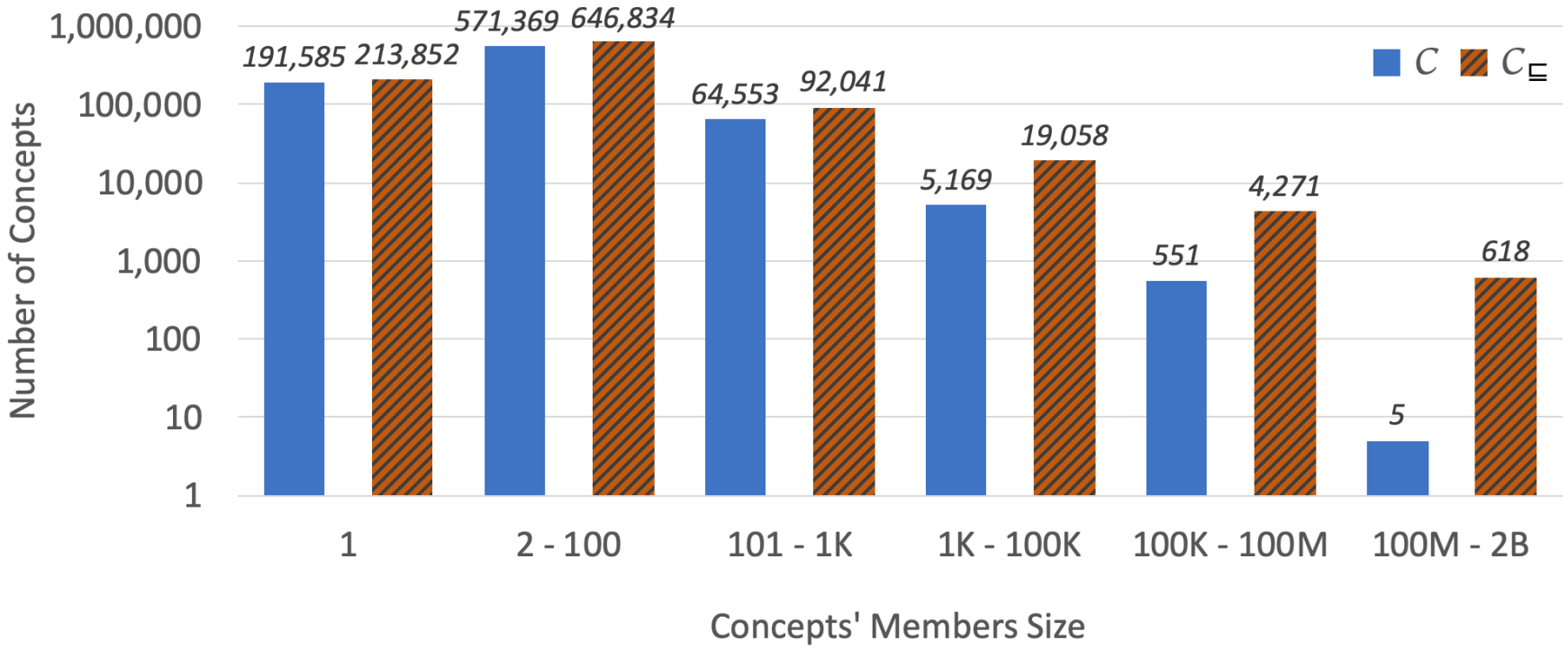


CommonKADS : 3rd KADS Meeting : Siemens AG Munich, March 8 - 9,1993 ; organized with GI special interest group 1.5.1 Knowledge Engineering

Dataset

# triples	28,362,198,927
# rdf:type statements	3,321,354,308
# rdfs:subClassOf statements	4,461,717
# owl:equivalentClass statements	1,051,979
# explicit owl:sameAs statements	558,943,116
# implicit owl:sameAs statements	35,201,120,188
# equivalence classes (after closure of owl:sameAs)	48,999,148
# concepts with at least one explicit member $\|C\|$	833,232
# concepts with at least one explicit or implicit member $\|C_{\sqsubseteq}\|$	976,674

Size distribution of the Concepts' members



- **23%** of the concepts have one explicit member
- **92%** of the concepts have ≤ 100 explicit members
- **618 concepts** have more than 100M explicit or implicit members
- **5 concepts** have more than 100M explicit members

Concepts with more than >100M explicit members

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/ sensorobservation.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**

These **5 concepts** with more than 100M explicit members
are the objects of 62% of the total rdf:type statements in the LOD-a-lot

Experiments

Research Question

Does the inclusion of instance-level interlinks (i.e. owl:sameAs) positively impact instance-based schema alignments ?

a. Does the inclusion of owl:sameAs increase the Jaccard Index of equivalent concepts?

b. Does the inclusion of owl:sameAs increase the Jaccard Index of non-equivalent concepts?

Dataset

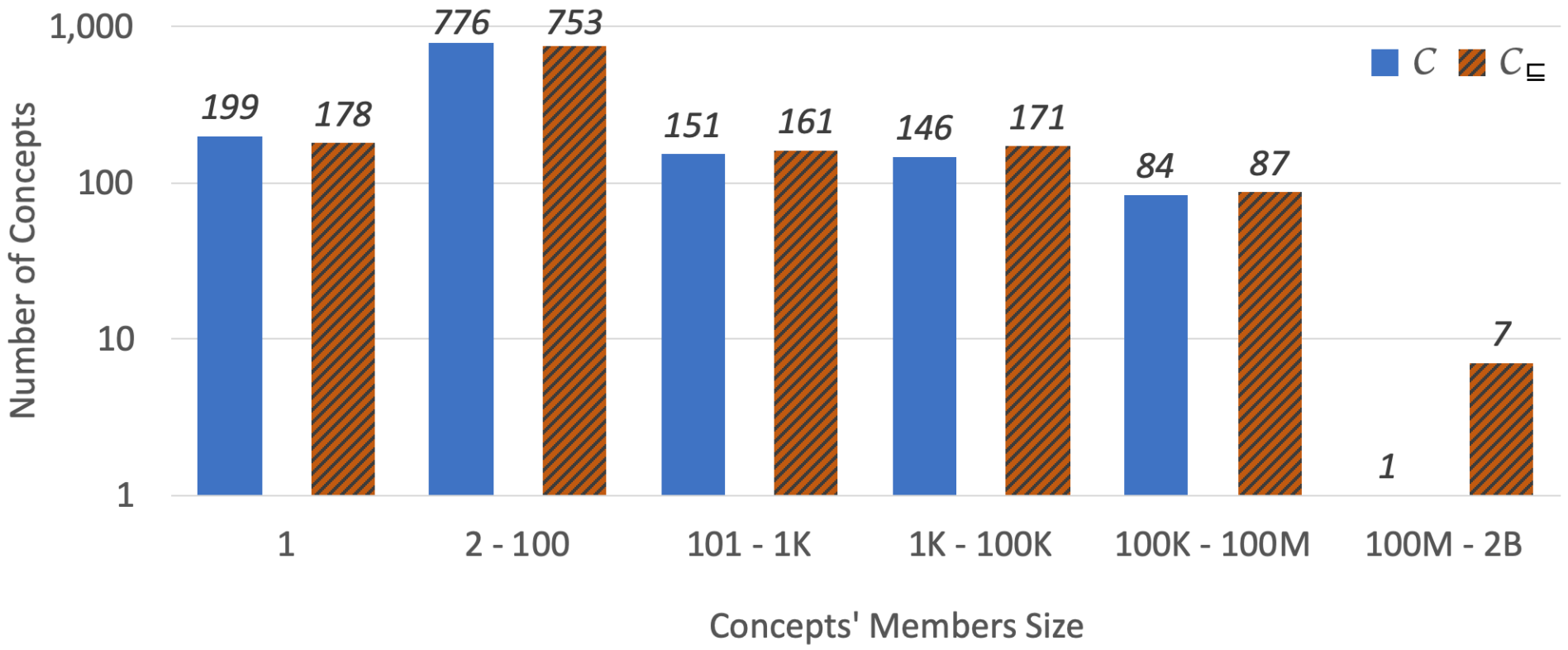
# triples	28,362,198,927
# rdf:type statements	3,321,354,308
# rdfs:subClassOf statements	4,461,717
# owl:equivalentClass statements	1,051,979
# explicit owl:sameAs statements	558,943,116
# implicit owl:sameAs statements	35,201,120,188
# equivalence classes (after closure of owl:sameAs)	48,999,148
# concepts with at least one explicit member $\|C\|$	833,232
# concepts with at least one explicit or implicit member $\|C_{\sqsubseteq}\|$	976,674

a. Does the inclusion of owl:sameAs increase the Jaccard Index of equivalent concepts?

- 1,051,979 owl:equivalentClass statements in the LOD-a-lot
  - Hypothesis: all these existing statements are correct alignments
  - Only 972 owl:equivalentClass statements where both concepts have explicit members
    - 208 reflexive alignments *(C1, owl:equivalentClass, C1)*
    - 22 duplicate symmetric alignments *(C1, owl:equivalentClass, C2) and (C2, owl:equivalentClass, C1)*
      - **742 alignments between 1,357 distinct concepts (i.e. gold standard)**

a. Does the inclusion of owl:sameAs increase the Jaccard Index of equivalent concepts?

Size distribution of the Concepts' members of our Gold Standard



a. Does the inclusion of owl:sameAs increase the Jaccard Index of equivalent concepts?

Jaccard Index distribution for the 742 alignments

Compare

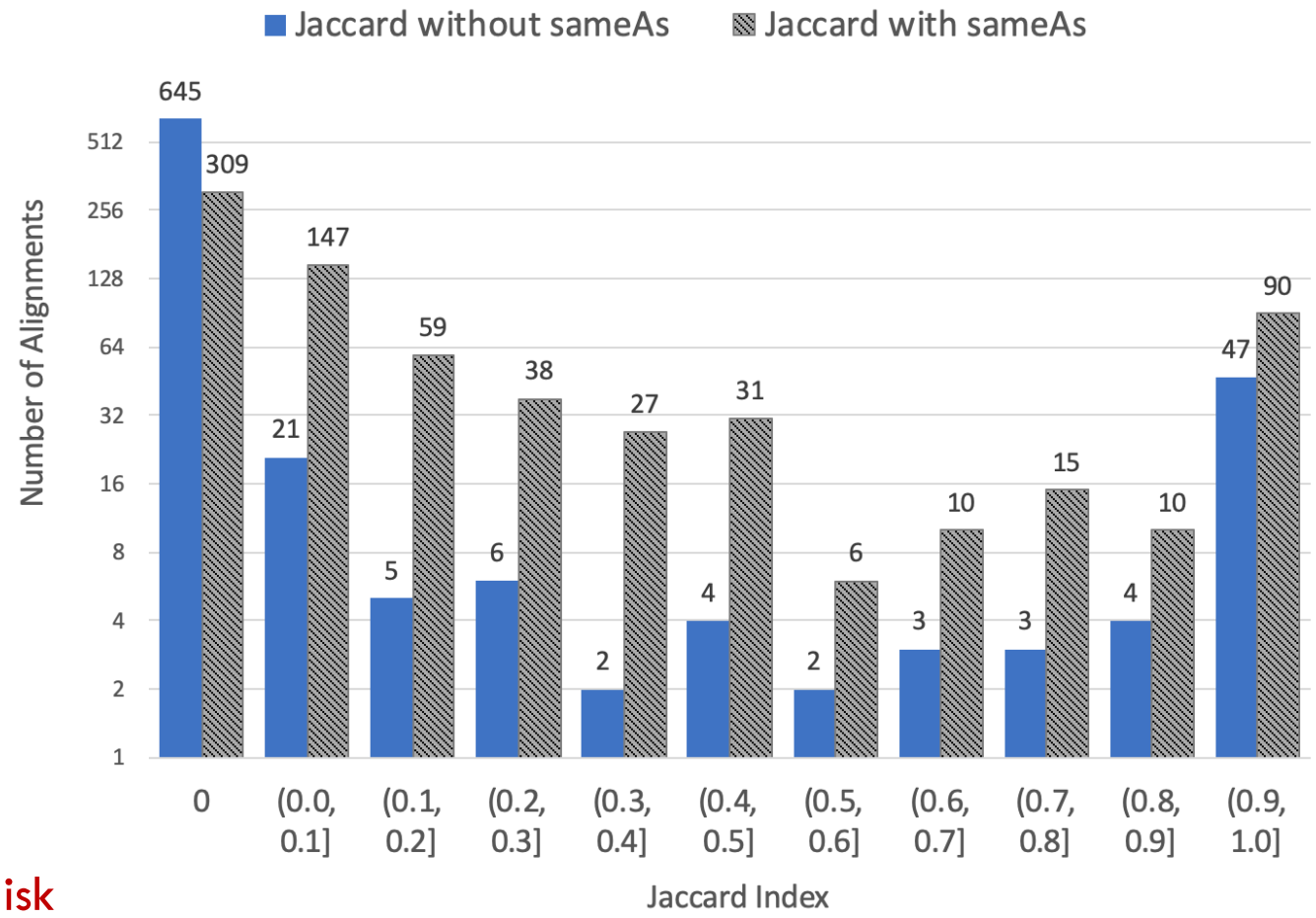
J(ext$_{\sqsubseteq}$(C1), ext$_{\sqsubseteq}$(C2))

with

J(ext$_{\sqsubseteq}$~(C1), ext$_{\sqsubseteq}$~(C2))

such that

(C1, owl:equivalentClass, C2)



Runtime: 4 hours on 64GB SSD disk

a. Does the inclusion of owl:sameAs increase the Jaccard Index of equivalent concepts?

Jaccard Index variation for the 742 alignments

- When owl:sameAs is considered, the Jaccard index increases for 381/ 742 of the correct alignments (52%)
- Out of these 381 cases, Jaccard increases from 0 to 1 in 44 cases (6%)
- When owl:sameAs is considered, the Jaccard index decreases for 25/ 742 of the correct alignments (3%)
- Slight drop in impact when only explicit members are considered

The inclusion of owl:sameAs does increase the overlap of two equivalent concepts in half of the cases

	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%)***

Research Question

Does the inclusion of instance-level interlinks (i.e. owl:sameAs) positively impact instance-based schema alignments ?

a. Does the inclusion of owl:sameAs increase the Jaccard Index of equivalent concepts?

b. Does the inclusion of owl:sameAs increase the Jaccard Index of non-equivalent concepts?

Dataset

# triples	28,362,198,927
# rdf:type statements	3,321,354,308
# rdfs:subClassOf statements	4,461,717
# owl:equivalentClass statements	1,051,979
# explicit owl:sameAs statements	558,943,116
# implicit owl:sameAs statements	35,201,120,188
# equivalence classes (after closure of owl:sameAs)	48,999,148
# concepts with explicit members $\|C\|$	833,232
# concepts with explicit or implicit members $\|C_{\sqsubseteq}\|$	976,674

b. Does the inclusion of owl:sameAs increase the Jaccard Index of non-equivalent concepts?

- 833,232 concepts with explicit members in the LOD-a-lot
  - Create one random alignment for each concept, such that each concept is paired only once
  - Hypothesis: all these random alignments are erroneous
    - **416,616 random alignments**

b. Does the inclusion of owl:sameAs increase the Jaccard Index of non-equivalent concepts?

Jaccard Index variation for the 416,616 random alignments

- When owl:sameAs is considered, the Jaccard index increases for only 94 / 416,616 of the random alignments (0.02%)
- When owl:sameAs is considered, the Jaccard index decreases for 3 / 416,616 of the random alignments

owl:sameAs rarely increases the overlap of two non-equivalent concepts

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%)***

Take away message

Take away message

This work provides an empirical study on the impact of including instance-level interlinks on the overlap between concepts members

- Including instance-level interlinks can enhance the performance of instance-based schema alignments
  - Increases the overlap for 52% of the existing (i.e. correct) alignments in the LOD-a-lot
  - Increases the overlap for less than 0.3% of randomly created (i.e. erroneous) alignments

- Inference does positively impact instance-based schema alignments
  - Considering also the implicit members enhances the results on the Gold Standard by 3 pp

Additional findings in the paper:

- Discarding only isolated owl:sameAs links in the network can increase the quality of instance-based schema alignments (owl:sameAs links are probably not as bad as we first thought)
  - Reduces the cases where Jaccard index increases for non-equivalent concepts by 71%

On the Impact of sameAs on Schema Matching

JOE RAAD ◉ ERMAN ACAR ◉ STEFAN SCHLOBACH

VU VRIJE UNIVERSITEIT AMSTERDAM

Knowledge Representation & Reasoning Group

Thank you for your attention!

Code & Results

https://github.com/raadjoe/impact-sameAs-schema-matching