

# Refining Large Identity Graphs using the Unique Name Assumption

Under submission at ESWC 2022 (research track)

Shuai Wang, Joe Raad, Peter Bloem, Frank van  
Harmelen

KR&R Group, Vrije Universiteit Amsterdam

10th January, 2022

A decorative geometric pattern in the bottom right corner of the slide, consisting of a grid of triangles in various shades of blue, arranged in a way that creates a sense of depth and movement.

# Content

1. Introduction
2. Testing UNA
3. Algorithm
4. Evaluation
5. Contributions
6. Discussion

# Introduction

## Introduction

### Testing UNA

The Gold Standard  
Validating UNA  
Reliability

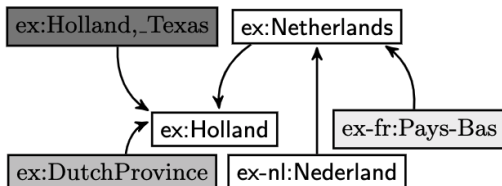
### Algorithm

### Evaluation

### Contributions

### Discussion

- (A, owl:sameAs, B).
- owl:sameAs is an equivalence relation : transitive, symmetric, reflexive.
- Error rate: 3 - 4%, or as high as 20%.



# Introduction

## Introduction

### Testing UNA

The Gold Standard

Validating UNA

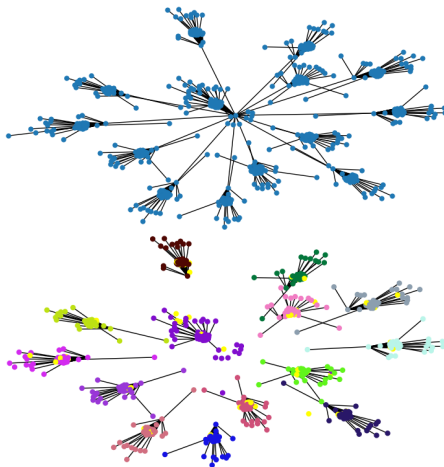
Reliability

### Algorithm

### Evaluation

### Contributions

### Discussion



# Related Work

## Introduction

### Testing UNA

The Gold Standard

Validating UNA

Reliability

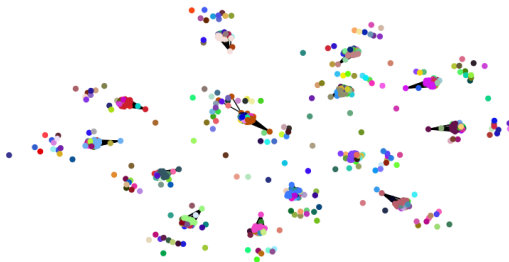
### Algorithm

### Evaluation

### Contributions

### Discussion

- content-based approach
- network-based approach (Louvain)
- **inconsistency-based approach**



# The UNA

Introduction

Testing UNA

The Gold Standard

Validating UNA

Reliability

Algorithm

Evaluation

Contributions

Discussion

- the naive UNA (nUNA): any two URIs in the same knowledge base cannot refer to the same thing in the real world.
- the quasi UNA (qUNA) extends this definition by taking the **redirect** relations (between 6 major hubs) and **dead** nodes into account.
- we also found entities that only differ in encoding: **encoding equivalence**.

# The Challenge

## Introduction

### Testing UNA

The Gold Standard

Validating UNA

Reliability

### Algorithm

### Evaluation

### Contributions

### Discussion

- need a large gold standard (so no reliable evaluation)
- no redirect graphs
- no graphs about encoding equivalence
- no definition about provenance
- no UNA definitions has been validated at large scale

# Provenance

## Introduction

## Testing UNA

The Gold Standard

Validating UNA

Reliability

## Algorithm

## Evaluation

## Contributions

## Discussion

let  $(e_i)$  be the sources of an entity  $e_i$ .

**Explicit sources:** the  $les$  where there are triples of `rdfs:isDefinedBy` \*.

**Implicit label-like sources:** the  $les$  where there are triples of `rdfs:label` \*.

**Implicit comment-like sources:** the  $les$  where there are triples of `rdfs:comment` \*.

\* or any equivalent relation or sub-properties



# internal UNA (iUNA)

## Introduction

### Testing UNA

The Gold Standard

Validating UNA

Reliability

### Algorithm

### Evaluation

### Contributions

### Discussion

iUNA: two different IRIs within the same [namespace](#) should refer to distinct real-world entities only when they are defined in the same source.

Exemptions:

- redirects
- encoding equivalence
- exceptions while resolving the IRI:
  - dead node
  - not found
  - unresolvable
  - redirects until reaching some error or not found
  - or has timeout error while resolving

# The Problem

## Introduction

Testing UNA  
The Gold Standard  
Validating UNA  
Reliability

## Algorithm

## Evaluation

## Contributions

## Discussion

**Issue** A very large subgraph about owl:sameAs (550 million nodes in LOD-a-lot).

**Task** remove as few edges as possible.

**Complexity** = APX-hard (has a polynomial-time approximation).

**Intuition** pull & push

# Research Questions

## Introduction

## Testing UNA

The Gold Standard

Validating UNA

Reliability

## Algorithm

## Evaluation

## Contributions

## Discussion

- RQ1: how can we define UNA for large integrated knowledge graphs?
- RQ2: how do we validate the definitions proposed?
- RQ3: can UNA give a reliable indication of identity errors in practise?
- RQ4: can we define an efficient algorithm for the refinement of the identity graphs?
- RQ5: is it possible to improve the results using additional information from the graph?

# The Gold Standard

Introduction

Testing UNA

The Gold Standard

Validating UNA

Reliability

Algorithm

Evaluation

Contributions

Discussion

- 8,394 manually annotated entities in 28 les
- a total of 232,311 owl:sameAslinks
- 11.75% entities are 'unknown'
- the error rate is between 1.58% and 9.98%

# Validating UNA

Introduction

Testing UNA

The Gold Standard

**Validating UNA**

Reliability

Algorithm

Evaluation

Contributions

Discussion

Does the UNA hold using label-like sources?

- nUNA: 93.50%
- qUNA: 94.43%
- iUNA: 94.11%

Using comment-like sources:

- nUNA: 97.46%
- qUNA: 96.77%
- iUNA: 97.09%

Yes, very much so!

Introduction

Testing UNA

The Gold Standard

Validating UNA

Reliability

Algorithm

Evaluation

Contributions

Discussion

**RQ3: Can the UNA give a reliable indication of identity errors in practice?**

**Baseline (label-like sources): error rate of random pairs: 47-68%.**

- **nUNA: 61.79% pairs violate; error: 33.31 - 49.89%.**
- **qUNA: 41.23% pairs violate; error: 33.28 - 51.87%.**
- **iUNA: 0.78% pairs violate; error: 6.10 - 36.69%.**

# Reliability: Redirect

Introduction

Testing UNA

The Gold Standard

Validating UNA

**Reliability**

Algorithm

Evaluation

Contributions

Discussion

Among them existing edges: error rate 1.47 - 7.69%.  
Others: error rate 4.29 - 6.32%.

# Reliability: Encoding Equivalence

Introduction

Testing UNA

The Gold Standard

Validating UNA

**Reliability**

Algorithm

Evaluation

Contributions

Discussion

Among them existing edges: error rate 2.21 - 8.50%.  
Others: error rate 1.16% and 14.83%.



# Algorithm

Introduction

Testing UNA

The Gold Standard

Validating UNA

Reliability

**Algorithm**

Evaluation

Contributions

Discussion

- 1 compute the minimum spanning forest (to reduce the load)
- 2 sample some edges from the original graph
- 3 assign an integer to each node
- 4 the (weighted) clauses are equivalence relations between these integers
- 5 positive weights for existing edges
- 6 negative weights for pairs violating the (chosen) UNA.
- 7 repeat until no such pair found or no edge removed

# Algorithm

Introduction

Testing UNA

The Gold Standard

Validating UNA

Reliability

**Algorithm**

Evaluation

Contributions

Discussion

# Algorithm: Weights

Introduction

Testing UNA

The Gold Standard

Validating UNA

Reliability

**Algorithm**

Evaluation

Contributions

Discussion

- the weighting scheme:  $w = (f_G; f_R; f_E; f_P)$
- the weight of an edge  $c_e$ :  
 $f_G(c_e) + f_R(c_e) + f_E(c_e) + f_P(c_e)$
- two weighting schemes for evaluation:  $w_1$  and  $w_2$

Introduction

Testing UNA

The Gold Standard

Validating UNA

Reliability

Algorithm

**Evaluation**

Contributions

Discussion

- Why precision-recall doesn't work anymore?
- a new measure

$$(G') = \sum_{C \in G'_{ccs}} \sum_{Q_e \in E(C)} \frac{|Q_e|}{|V|} \frac{|Q_e|}{|O_e|} \frac{|Q_e|}{|C|}.$$

- $C$  iterates over all connected components in  $G'$
- $E(C)$  is a partitioning of the nodes in  $C$  by equivalent class
- $O_e$  is the set of all entities in  $G'$  referring to  $e$ .

# Evaluation

Introduction

Testing UNA

The Gold Standard

Validating UNA

Reliability

Algorithm

**Evaluation**

Contributions

Discussion

	training set				evaluation set			
	precision	recall	$\Omega$	$ A $	precision	recall	$\Omega$	$ A $
Louvain algorithm	0.020	<b>0.759</b>	0.084	39,302	0.039	<b>0.660</b>	0.083	43,642
qUNA-label- $w_1$	0.300	0.061	0.587	<b>14</b>	<b>0.417</b>	0.006	0.607	57
qUNA-label- $w_2$	0.237	0.083	0.618	88	0.167	0.004	0.576	53
qUNA-comment- $w_1$	<b>0.324</b>	0.031	0.595	<b>14</b>	0.244	0.004	0.562	<b>24</b>
qUNA-comment- $w_2$	0.236	0.104	0.614	91	0.199	0.021	0.591	79
iUNA-label- $w_1$	0.186	0.077	0.605	101	0.086	0.026	0.585	35
iUNA-label- $w_2$	0.168	0.108	<b>0.619</b>	262	0.065	0.016	<b>0.617</b>	175
iUNA-comment- $w_1$	0.187	0.053	0.609	91	0.146	0.009	0.575	42
iUNA-comment- $w_2$	0.084	0.003	0.618	114	0.072	0.026	0.610	130

# Evaluation

Introduction

Testing UNA

The Gold Standard

Validating UNA

Reliability

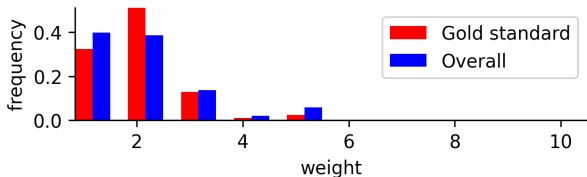
Algorithm

**Evaluation**

Contributions

Discussion

- weight: number of sources regarding each triple ( $0 < w < 650k$ ).
- DBpedia disambiguation nodes: corresponding to Wikipedia disambiguation pages.



# Evaluation: improving the results

Introduction

Testing UNA

The Gold Standard

Validating UNA

Reliability

Algorithm

**Evaluation**

Contributions

Discussion

	training set				evaluation set			
	precision	recall	$\Omega$	$ A $	precision	recall	$\Omega$	$ A $
iUNA-label- $w_2$	0.168	0.108	<b>0.619</b>	262	0.065	0.016	0.617	175
iUNA-label- $w_2$ +weight	0.217	0.108	0.610	233	0.050	0.015	0.614	162
iUNA-label- $w_2$ +disambiguation	0.221	0.135	0.615	264	0.098	<b>0.030</b>	<b>0.642</b>	191
qUNA-comment- $w_1$	0.324	0.031	0.595	<b>14</b>	<b>0.244</b>	0.004	0.562	<b>24</b>
qUNA-comment- $w_1$ +weight	0.159	0.016	0.579	17	0.111	0.002	0.575	27
qUNA-comment- $w_1$ +disambiguation	<b>0.412</b>	<b>0.163</b>	0.573	209	0.133	0.005	0.578	43

# List of Contributions

Introduction

Testing UNA

The Gold Standard

Validating UNA

Reliability

Algorithm

Evaluation

**Contributions**

Discussion

- 1 **UNA**: validated and checked the reliability of nUNA, qUNA, and iUNA.
- 2 **Algorithm** but does not scale to 177k.
- 3 **Datasets**: redirect, weights, encoding equivalence, disambiguation, etc.
- 4 **Gold standard**
- 5 **Results**: and how we improved it using additional information.



# Discussion

Introduction

Testing UNA

The Gold Standard

Validating UNA

Reliability

Algorithm

Evaluation

Contributions

Discussion

- 211,348 out of 232,311 edges (90.98%) are about DBpedia entities between different languages
- 177k nodes in the largest weakly connected component.
- only 5 have different label-like or comment-like sources: UNA is not about the source of errors.
- Next: evaluate using more methods
- Next: Deep Learning on the identity graph!

# Thank You for your attention!

Contact: [shuai.wang@vu.nl](mailto:shuai.wang@vu.nl)

