Examining the Evolution of Identity Links in the LOD Cloud*

Idries Nasim[1][0000−0001−8677−5218] and Shuai Wang[1][0000−0002−1261−9930]

Department of Computer Science, Vrije Universiteit Amsterdam, Amsterdam
idries.Nasim@gmail.com|shuai.wang@vu.nl

Keywords: semantic web evolution· identity graphs · knowledge graph

1 Introduction

The increasing adoption of the linked data principles brought challenges of large-scale data management, data provenance, and various quality-related issues. More and more linked data is being published which is in its best scenario[1] structured and interlinked data using Uniform Resource Identifiers (URIs). Thanks to the identity links between such data, one can retrieve more complete information from different sources [3]. However, as the web of linked data grows and evolves, the graphs corresponding to the identity links, the identity graphs, exhibit different properties. When a URI can no longer be dereferenced, some information associated with it can no longer be accessed, this results in not only the loss of information (label, comments) but also the errors of the software and knowledge systems depending on it. Another issue happens when an URI is redirected but the new entities are not linked or included in the knowledge bases. There can be confusion and mistakes for knowledge systems that rely on such identity links. This impairs the accessibility and reusability.

Some analysis indicates that around 19.4% entities in the identity graphs no longer exist after two years since its first publication [4]. To the best of our knowledge, the latest web-scale analysis of the identity links dates back to 2015 [2] consisting of 179 million unique URIs extracted from the 2015 LOD Laundromat crawl of the Web [1]. However, this study lacks statistical analysis on the accessibility of these entities. In this thesis, we present the construction of a new integrated identity graph (August 2022) and study the evolution by comparing against the old integrated identity graph (2015).[2]

2 Constructing the New Identity Graph

We perform an examination of the 643,886 datasets that are covered in the 2015 LOD Laudromat crawl [1]. Only 7,610 datasets (1%) contains `owl:sameAs`

[*] This is an extended abstract of the thesis supervised by the co-author, Shuai Wang.

[1] `https://5stardata.info/en/`.

[2] The code and data are available at `https://github.com/shuaiwangvu/scaling_identity_refinement`.

identity links. Our analysis shows that over 93.6% of identity links were published in 814 linksets (knowledge graphs that consist of only identity links) about 91.2% entities. Thus, the construction of the new identity graph targets mainly at updated linksets and major hubs, especially those that are well maintained. More specifially, we take only those published in or after 2018. We downloaded various linksets and data dumps to create our integrated identity graph including Yago4, Wikidata, Caligraph, IMDB, Wordnet, etc. The integrated graph has 409.3 million triples and 433.4 million distinct URIs stored in an HDT file of 11 GB. For easy reference, we name the old identity graph of sameas.cc G, and our newly constructed identity graph H.

3 Constructing the Redirection Graph

Two sampling were performed. Firstly, 100,000 entities were sampled uniformly from both graphs. To study if the size of connected component correlates to the number of hops of redirect, we sample entities from connected components (a maximal set of entities that are reachable to each other) of size 2, between 3 and 10, and larger than 10, denoted CC(2), CC(3-10), and CC(>10).

4 Analyzing the Evolution of Identity Graphs

There is an overlap of 57,884,691 unique entities between G and H. That is, 32.3% of the entities in the old identity graph still exist in comparison to previous estimate of 19.4% [4]. These entities count only 13.4% of our new identity graph, which indicate that the entities in the identity graphs of the LOD cloud have changed significantly. Table 1 shows statistics of entities in the redirect graph that face 400+ HTTP error indicating a client error and are not found (NF), found without redirect (OK), not found with errors (ER), not found due to timeout (TO), redirected until found (RUF), redirected until not found (RUNF) and that of timeout (RUT). For H, more entities were found after redirect and less faces timeout. In addition, when uniformly sampled, our results show that both graphs have an average step of redirect of around 2.21 hops. More entities in G were not found (67,03%). Finally, we found that only few of the redirected URIs of G also occur in graph H. In conclusion, our analysis shows that redirection can be used for the discovery of new URIs but not used as identity links, nor to bring existing mappings to date.

Table 1: results of redirect analysis

	Sampling	**#Entities**	**NF (%)**	**OK(%)**	**ER(%)**	**TO(%)**	**RUT(%)**	**RUNF(%)**	**RUE(%)**	**RUF(%)**
G	uniform	100,000	13.3	1.1	23.9	8.2	8.1	12.8	<0.1	32.6
	CC(2)	20,000	4.0	0.7	39.5	12.3	0.9	5.5	0.0	37.1
	CC(3-10)	20,000	8.4	0.3	43.4	5.8	0.9	5.8	5.0	30.4
	CC(>10)	20,000	11.0	0.8	26.5	23.2	2.3	10.1	0.1	26.0
H	uniform	100,000	14.8	3.0	18.5	2.4	4.5	12.3	0.1	44.4
	CC(2)	20,000	15.8	6.7	3.2	3.4	8.0	8.4	<0.1	54.5
	CC(3-10)	20,000	6.9	2.4	59.1	4.9	2.5	6.2	0.1	17.9
	CC(>10)	20,000	3.4	4.1	67.8	3.9	2.3	4.4	<0.1	14.1

References

1. Beek, W., Rietveld, L., Bazoobandi, H.R., Wielemaker, J., Schlobach, S.: Lod laundromat: a uniform way of publishing other people's dirty data. In: ISWC. pp. 213–228. Springer (2014)
2. Beek, W., Raad, J., Wielemaker, J., van Harmelen, F.: sameas.cc: The closure of 500m owl:sameas statements. In: ESWC. Springer (2018)
3. Debattista, J., Lange, C., Auer, S., Cortis, D.: Evaluating the quality of the lod cloud: An empirical investigation. Semantic Web **9**, 1–43 (08 2018). https://doi.org/10.3233/SW-180306
4. de Melo, G.: Not quite the same: Identity constraints for the web of linked data. In: AAAI (2013)