

What does it mean when your URIs are redirected? Examining identity and redirection in the LOD cloud

Idries Nasim¹, Shuai Wang^{1,*}, Joe Raad², Peter Bloem¹ and Frank van Harmelen¹

¹*Department of Computer Science, Vrije Universiteit Amsterdam, Boelelaan 1111, Amsterdam, the Netherlands*

²*Interdisciplinary Laboratory of Numerical Sciences (LISN), University of Paris-Saclay, Orsay, France*

Abstract

Redirection of URIs is widely used in the LOD cloud, and is even part of the best practice guidelines as an approach to the “curation problem” on the semantic web (i.e. how to repair imperfections). When dereferencing, one URI is redirected to another URI. Such a redirection could be the result of an update of the namespace, a different encoding scheme, or some other reasons. In this paper, we study the semantics of redirection and examine if redirection indicates how entities in the LOD cloud evolve. More specifically, we focus on entities in the identity graphs: subgraphs in the semantic web restricted to identity links. The entities we study are from `sameAs.cc`, an identity graph extracted from a crawl of the semantic web in 2015. Our analytical results include an examination of edges and chains of redirection as well as a statistical analysis of the redirection behavior of sampled entities. Additionally, we present properties of the graphs formed by redirection relations.

Keywords

identity graphs, knowledge graph evolution, semantic web evolution, identity crisis, redirection

1. Introduction

The semantic web is a decentralised world-wide information space for sharing machine-readable data about entities and their relations. This information space contains a vast and rapidly increasing quantity of scientific, corporate, government, and crowd-sourced data openly published on the Web. Open Data plays a catalyst role in the way structured information is exploited on a large scale. In this space, resources are identified by global identifiers called Uniform Resource Identifiers (URI). A traditional view of digitally preserving these resources is by “pickling and locking them away” for future use, like groceries, but this conflicts with their evolution. Instead, when resources change or become outdated, a common (and even recommended) solution to the “curation problem” (i.e. repairing data imperfections) is to redirect the user or agent to a new location. We investigate how such redirections can indicate the evolution of entities in the cloud of linked open data.

Managing the Evolution and Preservation of the Data Web (MEPDaW 2022)

*Corresponding author.


✉ m.i.nasim@student.vu.nl (I. Nasim); shuai.wang@vu.nl (S. Wang); joe.raad@lisn.fr (J. Raad); p.bloem@vu.nl (P. Bloem); frank.van.harmelen@vu.nl (F. v. Harmelen)

🌐 <https://shuai.ai/research> (S. Wang); <http://www.joe-raad.com/> (J. Raad); <https://peterbloem.nl/> (P. Bloem); <https://www.cs.vu.nl/~frankh/> (F. v. Harmelen)

🆔 0000-0001-8677-5218 (I. Nasim); 0000-0002-1261-9930 (S. Wang); 0000-0002-7891-7738 (J. Raad); 0000-0002-0189-5817 (P. Bloem); 0000-0002-7913-0048 (F. v. Harmelen)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

Semantic web resources can be divided into two main categories¹: information resources whose essential characteristics can be conveyed in a message (e.g. web pages, documents), and non-information resources that are outside the information space of the Web (e.g. Amsterdam, Tim Berners-Lee, the concept of color). When dereferencing an outdated URI of a non-information resource such as the city of Amsterdam (e.g. <https://dbpedia.org/resource/Amsterdam>), it is best practice [1] to redirect the user or agent to the information resource about this city (e.g. <https://dbpedia.org/page/Amsterdam>) using the HTTP response code 303 known as ‘see other’.

In practice, redirections through 3XX response codes are not limited to such cases, and are as well used to prevent information loss when a URI can no longer be dereferenced. Precisely, redirecting between two information resources (e.g. in case of a website’s update) or between two non-information resources (e.g. for preserving backwards compatibility when an RDF dataset is updated). As the semantic web develops, such redirection links capture the information evolution between URIs. In fact, when dereferencing a URI there can be multiple intermediate URIs involved in the redirection. For instance, Figure 1 illustrates different scenarios that occur in practice when dereferencing URIs². It shows five entities of an RDF graph: e_0 , e_3 , e_6 , e_8 , and e_9 that are connected by any object property, represented in this figure with the black edges (in this paper we will restrict to `owl:sameAs` identity links). Red edges represent HTTP redirection links, showing for instance a redirection from e_3 to e_5 with an intermediate redirection to e_4 . The links from e_0 are an example of redirections that ultimately lead to an error (e.g. because of an 4XX response code when dereferencing e_2), illustrated as a cross-out node. Finally, this figure shows another case where two resources (e_6 and e_9) are redirected to the same URI (e_7), before reaching e_{10} , which faces a timeout error (denoted using a question mark) when attempting to resolve the URI it redirects to.

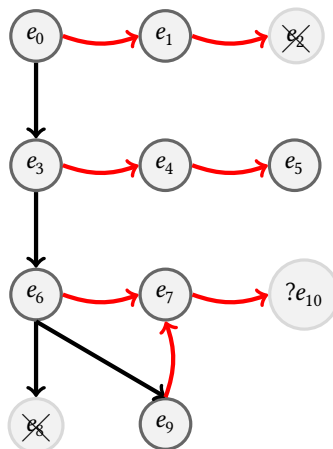


Figure 1: An illustration of HTTP GET request of URIs (black arrows for `owl:sameAs` and red arrows for redirection)

¹See <https://www.w3.org/2001/tag/doc/httpRange-14/2007-05-31/HttpRange-14> for more details about URIs, dereferencing, redirection, (non-)information resources, and their relations.

²See Section 3.2 and Table 1 for our annotation of different scenarios.

Although these redirection mechanisms are an integral part of the architecture of the web, and are part of the best practice guidelines for linked data, the semantics of such redirection is unclear. It is tempting to identify a redirection with an implicit statement of identity: the source of the redirection is semantically equivalent to the target of the redirection, it is only the location of the resource is different. In this paper we set out to clarify the semantic intent of redirections as they are being used in practice.

We investigate how redirection can indicate the evolution of entities in the cloud of linked open data (i.e. the LOD cloud). More specifically, we focus on redirection of entities in subgraphs that are restricted to identity links between entities (such as `owl:sameAs`). When considering the URIs identical to entities in these subgraphs, there can be ambiguity and unwanted consequences due to the semantics of identity links. We study the following two research questions.

RQ1: Can we approximate the implicit semantics of redirection?

For this question, we examine sampled edges and chains of redirection. We classify the scenarios of redirection and estimate the proportion of redirection that can be interpreted as identity links.

RQ2: What are the properties and structure of the redirection graphs?

To answer this question, we study the redirection graphs by performing a statistical analysis and examining their graph-theoretical properties, followed by a discussion about its impact on the LOD cloud.

Our main contributions are as follows:³

1. four redirection graphs corresponding to different sampling methods using the `sameas.cc` identity graph;
2. 4,000 semi-automatically annotated edges (as pairs of URIs) in the uniformly sampled redirection graph;
3. a qualitative study of the semantics of redirection in the identity graphs;
4. a quantitative study of properties of the redirection graphs.

The paper is organised as follows. In Section 2, we present related work on redirection and identity graphs. Section 3 introduces the new redirection graphs, based on which we sample data for analysis. Section 4 studies the semantics of redirection. The analysis of the redirection graphs is discussed in Section 5 followed by conclusions and future work in Section 6.

2. Related Work

As a domain with a strong focus on unambiguous identifiers and meaning, semantic web research has been suffering from an ill-defined sense of identity [2]. This crisis becomes even worse when taking into account the impact of the evolution of datasets on the identity links. The identity crisis was already studied by Halpin, et al. [3] in 2009. They propose to study how an HTTP resource responds to a GET request, including how they redirect to new URIs (both the hash convention and the HTTP 303 redirection). However, this work did not retrieve or

³The source code can be found at <https://github.com/shuaiwangvu/redirection>. The datasets were published online at <https://doi.org/10.5281/zenodo.7225383> with DOI 10.5281/zenodo.7225383.

study any data from the web, nor performed any quantitative assessment on the reliability of interpreting redirection as identity relations.

The evolution of datasets can result in missing URIs. De Melo presented an initial analysis in 2013 and revealed that, for the BTC2011 sameAs triples, 205,231 out of 1,055,626 unique DBpedia URIs did not exist in the DBpedia 3.7 dataset [4]. This analysis shows that around 19.4% of entities no longer existed after only two years since their first publication. The paper also investigated the reasons for this. For example, URIs with incorrectly escaped titles, i.e. using a different encoding scheme than DBpedia itself, resulted in URIs that do not exist in DBpedia. Secondly, since Wikipedia is a living resource, articles may be deleted, merged, or renamed. Thus, many URIs no longer exist in DBpedia.

Regino et al. [5] studied semantically broken links. These are newly added links between the new URIs of the subjects or objects that may have evolved. When the evolved URIs refer to different real-world entities, the change of semantics would result in errors (thus the name “semantically broken links”). For example, a link between e_3 and e_4 in Figure 1 could be such an example if e_5 refers to a different real-world entity. They studied the links between Wikidata and GeoNames and two versions of DBpedia. While their analysis found some semantically broken links, their approach cannot be scaled to the web since they only studied English entities and rely on WordNet and BabelNet as background knowledge for the determination of similarity by analyzing on their labels. Moreover, tracking every version of entities in each dataset is not practically feasible.

To the best of our knowledge, the latest web scale examination of the identity graphs dates back to the 2015 crawl of the web⁴. It consists of 558.9M owl:sameAs links between about 179.7M entities [6]. However, this graph is now outdated, and as far as the authors are aware, there is no quality assessment of its entities, in comparison to the presence of multiple assessment of its links. In contrast, the current paper aims at addressing the importance of dynamics in identity graphs.

3. Data Preparation

In this paper, we extract our entities from the sameas.cc dataset [6]. This identity graph represents a subgraph restricted to owl:sameAs links of the 2015 LOD Laundromat dataset [7] that covers more than 650K datasets. We refer to this identity graph as G . Section 3.1 provides details of sampling. Based on the sampled entities, we construct the redirection graphs in Section 3.2. Finally, in Section 3.3 we sample 4,000 edges and 100 chains of redirection in the redirection graph based on uniformly sampled entities. These datasets will be analyzed in Section 4 and 5 to answer our research questions.

3.1. Sampling from identity graphs

For this study, four samples were created. The first sample E^U is created by randomly choosing 100K entities from G . The remaining three samples contains 20K entities each, with the goal of studying the presence of a correlation between the size of the connected components of G (CC)

⁴The resulting identity graph and its related research results are hosted at <https://sameas.cc>.

and the semantics of redirection. In G , the set of entities in a CC refer to an equivalence class (i.e. set of entities that refer to the same real-world entity). These entities were sampled equally from CCs containing only 2 entities, ones containing 3 to 10 entities, and CCs with more than 10 entities. We refer to these samples as $E^{CC(2)}$, $E^{CC(3-10)}$ and $E^{CC(>10)}$, respectively.

3.2. Constructing the redirection graphs

We analyse the URI of the sampled entities by sending an HTTP GET request. If the response status code is HTTP 200, we label it as **OK**. If it is a 400+ HTTP error indicating a client error, we label it as ‘Not Found’ (**NF**). Otherwise, if the entity is a literal or the request fails, we label it with ‘Error’ (**ER**). We use the label ‘Timeout’ (**TO**) if the request times out. In practice, some URI takes longer to connect or read. Hence, we increase the timeout threshold in three steps. We first set the connection timeout to 0.01 second and read timeout parameters to 0.05 second. We collect all URIs with a timeout for processing in the next step and add labels to the rest. We then use the parameters 0.5 and 2.5 seconds and again collect those that faced a timeout. Finally, our last attempt uses 5 and 25 seconds as parameters. As for cases with redirection we used the history in the response to check if redirection happens. Thus, we include also HTTP 300 (redirection with multiple choice), 301 (moved permanently), 307 (temporal redirect), 308 (permanent redirect), etc. We label the remaining as ‘Redirect Until Timeout’ (**RUT**). Similar as above, we label URIs that redirect as either ‘Redirect Until Not Found’ (**RUNF**), ‘Redirect Until Error’ (**RUE**), or ‘Redirect Until Found’ (**RUF**). We create an edge in the redirection graph for each redirection. Similarly to the uniform sampling, we name this graph R^U for G , and similarly we name the three redirection graphs $R^{CC(2)}$, $R^{CC(3-10)}$, and $R^{CC(>10)}$ corresponding to the sampled entities $E^{CC(2)}$, $E^{CC(3-10)}$, and $E^{CC(>10)}$, respectively.

All the scripts were written in Python⁵. We performed all the HTTP GET requests on a cluster on August 23, 2022. The cluster has 32 CPUs of Intel Xeon E5-2630 v3 (2.40GHz) with 256GB of memory running Ubuntu 18.04.6. Its downloading speed is 871.56 MB/s. The construction of the redirection graphs took 33.5 hours in total.

3.3. Sampling edges and chains for manual analysis

To understand what these redirections are about, we sampled 4,000 edges from R^U . These edges are stored in a file as pairs of URIs. Moreover, we track the redirection behavior of 100 entities whose number of hops of redirection is greater than two. These chains will then be manually analyzed in the next section.

4. Implicit Semantics of Redirection

Next, we estimate the implicit semantics of each of these redirections (RQ1). In this section, we perform a qualitative analysis of redirection in the identity graphs. More specifically, Section 4.1 studies pairs of redirection and Section 4.2 provides details of our manual assessment of chains of redirection.

⁵All the code and scripts are open source in the repository at <https://github.com/shuaiwangyu/redirection>.

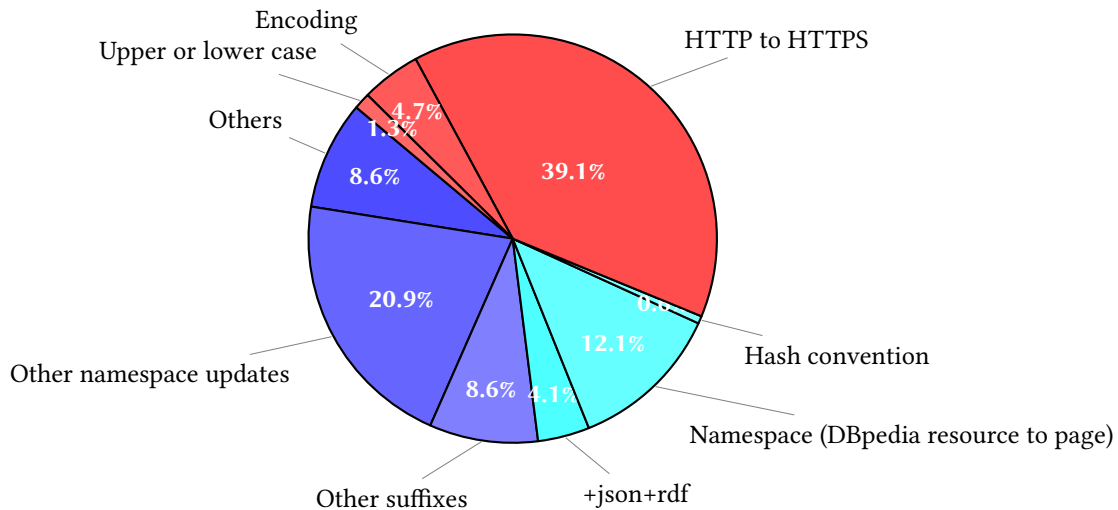


Figure 2: Proportion of redirection behavior among sampled entities

4.1. Analysing pairs of redirection

In this section, we study the nature of redirection links. For this, we sample 4,000 redirection links from R^U for semi-automatic analysis. Figure 2 illustrates the proportion of different cases. We found that 39.1% of URIs in R^U redirect to their https equivalent. A further 4.7% of URIs only differ from their redirect by encoding, while another 1.3% in R^U redirects to URIs only differ in upper/lower case. Together, this amounts to 45.1% of redirects that are mainly concerned with engineering technicalities. A second very common case are the updates of namespaces in the same domain (33.0%). For example, <https://www.worldcat.org/oclc/67950327> redirects to <https://www.worldcat.org/title/pro-patria/oclc/67950327>. We surmise that these are the result of dataset evolution. A specific case of intra-namespace redirections are the 12.1% that redirects from a DBpedia resource to a DBpedia page (but not the inverse). For example, https://dbpedia.org/resource/Rimula_californiana to http://dbpedia.org/page/Rimula_californiana. These are redirects from a representation to a description⁶. In addition, we found some cases where some suffixes are added to the original URIs (12.7%), including ‘.json’ (4.0%) and ‘.rdf’ (0.1%). Another 0.6% is about automatic truncation of fragment of hash URIs (i.e. the hash convention). Finally, various other cases make up the remaining 8.6%, with new URIs updated with ids and names, embedded queries, mistaken encoding, or other complex cases.

Our analysis shows that the HTTPS protocol has been widely adopted over the past years. It is likely that the semantics are preserved if they only differ by the choice of protocol. Similarly, if two URIs only differ by encoding or upper/lower case in their names, they are also likely to refer to the same real-life entities. This sums up to 45.1% (colored red, indicating identity preserving). As for redirection from non-information resources to information resources, only less than 1% concerns hash convention. We also observed at least 4.1% redirects to its corresponding files (with suffix of .json or .rdf) or from DBpedia resources to the corresponding pages (12.1%). This

⁶In the sense of <https://www.w3.org/TR/cooluris>.

sums up to 16.8% (colored cyan, indicating non-identity preserving). Given all the results, our best approximation is that between 45.1% and 83.2% (100%-16.8%) of redirection links can indeed be taken as identity links.

This primitive analysis shows that the semantics of redirection is rich in practice and requires further investigation with more detailed semi-automatic analysis. Given our analysis that a sizeable share of redirects (up to over half of them) cannot be reliably assumed to signify an identity link, we conclude that redirection should not be used to update outdated mappings without further refinement or manual assessment.

4.2. Analysing chains of redirection

Next, we perform an analysis of chains of redirection in R^U . On average, redirection chains have 1.7 hops. More precisely, entities redirected before timeout (RUT) take on average 1.7 hops to reach. Those redirected until not found (RUNF) take 1.6 hops. Those redirected until found (RUF) take 1.8 hops on average. Finally, there are only few redirected until error (RUE) with an average of 1.5 hops. Given the little difference we observed between each category, we uniformly sample 100 chains of redirection across these categories.

We extract 100 chains of redirection with at least 2 hops. Our manual examination shows that the individual redirections in these chains are rarely restricted to a specific type (from Section 4.1) but rather mix multiple types. This makes it very difficult to classify these chains. We also observe that these redirections mostly happen within a domain (85%). Among these chains, redirects within the domain `wikidata.org` is most common (28%). Redirection between DBpedia’s resources, pages, and their various encodings are also very common (26%). Moreover, these chains are among the longest in our sample with an average number of hops of 3.2. Other domains that occur frequently in these chains are `bibsonomy.org` (5%) and `viaf.org` (1%).

Table 1: Behavior of HTTP GET request of entities

Graph	RUF	OK	Valid ¹	ER	TO	RUT	RUNF	RUE	NF	Invalid ²
R^U	32.6%	1.1%	33.7%	23.9%	8.2%	8.1%	12.8%	0.01%	13.3%	66.3%
$R^{CC(2)}$	37.1%	0.7%	37.8%	39.5%	12.3%	0.9%	5.5%	0.0%	4.0%	62.2%
$R^{CC(3-10)}$	30.4%	0.3%	30.7%	43.4%	5.8%	0.9%	5.8%	5.0%	8.4%	69.3%
$R^{CC(>10)}$	26.0%	0.8%	26.8%	26.5%	23.2%	2.3%	10.1%	0.1%	11.0%	73.2%

¹ The valid entities include RUF (redirected until found), OK (found with HTTP 200)

² The rest are invalid entities, including ER (error), TO (timeout), RUT (redirected until timeout), RUNF (redirected until not found), RUE (redirected until error), and NF (not found).

5. Analyzing the Redirection Graphs

Table 1 shows an analysis of the behavior of HTTP GET request when applying our redirection typology to G (see Section 3.2 for the name of each column). When sampled uniformly, only 33.7% of the URIs are valid entities: information of the URI can be found (HTTP 200) with or without redirection (i.e. the sum of the ‘OK’ and ‘RUF’ column).⁷ Surprisingly, this result implies that only around 1% of the URIs return meaningful information directly. A comparison

of the column ‘OK’ with ‘RUF’ shows that redirection is a well adapted means to provide updated information for outdated URIs. In contrast, a disappointing 66.3% of entities are invalid: URIs that led to an error, could not be found, or resulted in a timeout (even after a few hops of redirection). When examining sampling w.r.t. connected components (CCs) of different sizes, we observe that the proportion of valid URIs decreases as the size of the CC increases. Correspondingly, the opposite trend shows for columns labelled ‘NF’ (not found), ‘TO’ (time out), ‘RUNF’ (redirect until not found), or ‘RUE’ (error). This would suggest that large connected components are a signature of poorly maintained subsets of URIs. This could be associated to the greater proportion of invalid entities as the size of CC increases. This might provide a useful heuristic for LOD maintenance.

Table 2 presents an analysis on how entities in E^U , $E^{CC(2)}$, $E^{CC(3-10)}$, and $E^{CC(>10)}$ are redirected. Over half of the entities are involved in redirection when sampled uniformly. The average hops of redirect is around 1.71. We observed that $R^{CC(3-10)}$ has a cycle of two entities redirecting to each other. The longest paths can be as many as 8 hops. Our manual examination shows that they are all about redirections between URIs involving DBpedia resources and pages.

Table 2
Properties of the redirection graph

Graph	#Entities	#Entities Redirected	#Nodes	#Edges	Avg #Hops	Max #Hops
R^U	100K	53,487 (53.49%)	169,021	116,031	1.71	8
$R^{CC(2)}$	20K	8,693 (43.46%)	30,091	21,602	1.64	8
$R^{CC(3-10)}$	20K	8,412 (42.06%)	29,697	21,490	-	-
$R^{CC(>10)}$	20K	7,704 (38.52%)	24,914	18,102	2.05	8

6. Conclusion

In this paper, we investigated different scenarios when URIs are redirected. We studied the semantics by examining edges and chains of redirection. The intuition behind redirects in the LOD cloud is that they preserve identity. Our analysis in section 4.1 shows that this is indeed the case for a large proportion of redirects sampled from the `sameas.cc` dataset, with 45% being almost certainly identity preserving, possibly up to 83%. In short, the answer to our first research question is that identity is indeed a plausible estimate of the semantics of redirects. However, given that for somewhere between 17-55% of redirects it is unclear whether they are identity preserving, we suggest that redirection should not be used to update outdated dataset mappings without further refinement or manual assessment.

In answer our second research question, concerning the properties and structure of the redirection graphs, we found that without any redirects, only 1% of all sampled URIs return meaningful information directly, rising to 33% after redirection. This means that a disappointing 66% of all URI’s end in error, failure or timeout at the end of their redirection chain. Furthermore, such failure cases are more frequent in larger connected components, suggesting that such

⁷As with the `sameas.cc` graph, we discovered a small number of literals. They were included as exceptions in the ‘ER’ column.

large connected identity components are indicative of poor maintenance, which may serve as a useful heuristic for LOD repair.

Section 4.1 presented an analysis of sampled redirection links. In future work, we would like to compare this distribution against existing identity links and study how similar they are. This could provide further evidence how we can take certain redirection links as identity links. Moreover, the identity graph we used is now considerably outdated. We could create a new updated identity graph and study redirects of sampled entities.

This paper restricted the analysis to entities in the identity graph. In future work, we would like to remove this restriction and compare against the redirection of URIs in the LOD cloud. Finally, it could be interesting to examine how redirection can help update existing mappings.

A possible use case could be to use a select a portion of redirection links for the refinement of identity graphs. Our analysis in Section 4.1 shows that 45.1% are considered identity-preserving. They could be used by refinement algorithms as additional information to improve the accuracy [8].

Finally, a reason that only around 1% of URIs still remain informative without redirection is that most URIs are managed by centralized registries, identity providers, and certificate authorities. Alternatively, Decentralized Identifiers (DIDs) [9] enable verifiable, decentralized digital identity. This could potentially be one of the means to resolve the issue of redirect for outdated URIs.

Acknowledgments

This project is a part of the MaestroGraph project, which is supported by the NWO TOP grant.

References

- [1] B. Hyland, G. Atemezing, B. Villazón-Terrazas, Best Practices for Publishing Linked Data, Technical Report, W3C Working Group, 2014. Online; accessed 19 October 2022.
- [2] R. Verborgh, M. Vander Sande, The Semantic Web identity crisis: in search of the trivialities that never were, *Semantic Web Journal* 11 (2020) 19–27. URL: <https://ruben.verborgh.org/articles/the-semantic-web-identity-crisis/>. doi:10.3233/SW-190372.
- [3] H. Halpin, V. Presutti, An ontology of resources: Solving the identity crisis, in: *ESWC 2009*, volume 5554, 2009, pp. 521–534. doi:10.1007/978-3-642-02121-3_39.
- [4] G. de Melo, Not quite the same: Identity constraints for the web of linked data, in: *AAAI*, 2013.
- [5] A. G. Regino, J. C. dos Reis, Discovering semantically broken links in LOD datasets, in: *Proceedings of the 6th Workshop on Managing the Evolution and Preservation of the Data Web (MEPDaW)*, 2020.
- [6] W. Beek, et al., sameas. cc: The closure of 500m owl: sameas statements, in: *European semantic web conference ESWC*, Springer, 2018, pp. 65–80.
- [7] W. Beek, et al., LOD laundromat: a uniform way of publishing other people’s dirty data, in: *ISWC*, Springer, 2014, pp. 213–228.

- [8] S. Wang, J. Raad, P. Bloem, F. van Harmelen, Refining transitive and pseudo-transitive relations at web scale, in: R. Verborgh, K. Hose, H. Paulheim, P.-A. Champin, M. Maleshkova, O. Corcho, P. Ristoski, M. Alam (Eds.), *The Semantic Web*, Springer International Publishing, Cham, 2021, pp. 249–264.
- [9] M. Sporny, D. Longley, M. Sabadello, D. Reed, O. Steele, C. Allen, *Decentralized Identifiers (DIDs) v1.0*, Technical Report, W3C, 2022. Online; accessed 19 October 2022.