The sameAs Problem: A Survey on Identity Management in the Web of Data

Joe Raad [a], Nathalie Pernelle [b], Fatiha Saïs [c], Wouter Beek [a] and Frank van Harmelen [a]

[a] *Dept. of Computer Science, Vrije Universiteit, Amsterdam, The Netherlands*
E-mails: j.raad@vu.nl, wouter@triply.cc, frank.van.harmelen@vu.nl
[b] *LIPN,Paris Sorbonne University, CNRS 7030, France*
E-mail: pernelle@lipn.univ-paris13.fr
[c] *LRI, Paris-Saclay University, CNRS 8623, Orsay, France*
E-mail: fatiha.sais@lri.fr

Abstract. In a decentralized global knowledge space such as the Web of Data, the `owl:sameAs` predicate is an essential ingredient. It allows parties to independently mint names, while at the same time ensuring that these parties are able to connect and complete each other's data. Since the manual creation of these links is expensive at large-scale contexts such as the Web of Data, identity links are often created automatically, with a chance of error. With several works already proven that identity in the Web of Data is broken, we investigate in this survey the approaches tackling this "sameAs problem", with a focus on (i) conducted studies and analyses of the identity use in the Web of Data, (ii) approaches proposing alternatives for `owl:sameAs`, (iii) approaches proposing identity management services, and (iv) ones focusing on detecting erroneous identity statements.

Keywords: Linked Open Data, Identity Analysis, Identity Management, Identity Invalidation, Contextual Identity

1. Introduction

In the era where the field of Artificial Intelligence (AI) is strongly dominated by Machine Learning, it is sometimes forgotten that the past decade has also seen a major breakthrough in Knowledge Representation (KR). Through the combination of web-technologies and a judicious choice of formal expressivity (description logics which correspond to a decidable 2-variable fragment of first order logic), it has become possible to construct and reason over knowledge graphs of sizes that were not realistic only few years ago. Nowadays, knowledge graphs of hundreds of millions of statements are routinely deployed by researchers from various fields (e.g. computer science, medicine, humanities), and companies worldwide (e.g. Google, Bing, Facebook). Since these knowledge graphs are mostly developed independently of one another, it is important that different organisations adhere to common principles and standards for encoding and publishing their knowledge. The most adopted set of principles were laid out by Tim Berners-Lee in 2010, and are known as the Linked Open Data (LOD) principles[1]. The idea is by providing simple best practices for creating structured data, publishers can also enrich, access, and benefit from a larger decentralised knowledge space, known as the Web of Data.

In such a large and distributed knowledge graph, it is common practice for the same real-world entity to be described in different knowledge graphs. In the absence of a central naming authority, it is also common practice for this same real-world entity to be denoted by different names (IRIs – Internationalised Resource Identifier). As such, Linked Open Data provides significant potential for knowledge interchange and reuse, since assertions about the same entity – possibly denoted by different names – may overlap and complement one another. In order to make the most of this available wealth of data, publishers are encouraged to link their data. Such interlinking is typically performed by asserting that two names

[1] https://www.w3.org/DesignIssues/LinkedData.html

in fact denote the same real-world entity. For this purpose, the Web Ontology Language (OWL) introduced the `owl:sameAs` predicate [1] that expresses the identity relation between resources. For instance, the RDF statement ⟨*Barack_Obama*, `owl:sameAs`, *44th_US_president*⟩ asserts that both names actually refer to the same person. Such identity statement also indicates that every property asserted to one name will be also inferred to the other, allowing both names to be used interchangeably in all contexts.

While such inferences can be extremely useful in enhancing a number of knowledge-based systems (e.g. providing more coverage and context for search engines, virtual assistants and recommendation systems), incorrect use of identity can have wide-ranging effects in a global knowledge space like the Web of Data. In fact, a number of studies over the years have already shown that identity is misused, estimating that around 3% [2] or 4% [3] of these links are erroneous, whilst others estimating this number to be in the range of 20% [4]. In addition, by exploiting the semantics of `owl:sameAs` and computing the transitive closure of over half a billion statements [5], we have showed the effects of such identity misuse in practice. Specifically, we showed that whilst in some cases identity misuse resulted in the false equivalence of semantically close entities (e.g. *Barack Obama* and the *Obama administration*), other cases have resulted in the false equivalence of over 177K names referring to a number of different countries, cities and people. With such findings leaving various uncertainties over the quality and usability of the Web of Data in its current state, proper approaches towards the handling of identity links are required in order to make the Web of Data succeed as an integrated knowledge space.

This survey provides the first literature review to this well-known "sameAs problem". It covers different families of works proposed for analysing and limiting this problem. First family of works focused on limiting the excessive use of `owl:sameAs`, by defining alternative identity relations that can replace `owl:sameAs` in certain contexts. Other initiatives tried limiting the misuse of `owl:sameAs` by developing centralised or decentralised services for identity link management and monitoring, such as sameas.org or sameas.cc. Finally, last family of works focused on (semi-) automatically detecting and flagging the incorrect `owl:sameAs` links. This survey presents and categorises these different solutions that aim to limit the identity problem in the Web of Data, and discusses their various strengths and drawbacks. As such, it does not cover related but distinct research topics like Entity Resolution and Reference Reconciliation, which focus on techniques and tools for establishing identity links (see [6] for a survey). This survey also does not address the historically significant, yet somewhat academic, distinction between *locating* an electronic document with a URL, and *denoting* an RDF resource with an IRI, known as the problem of Sense and Reference [7–9].

The rest of this paper is structured as follows. Section 2 gives an overview of the various aspects of the identity problem in the Web of Data, from a philosophical and practical point of view. Section 3 presents an overview of existing studies and analyses of the `owl:sameAs` usage in the Web. Section 4 presents current alternatives to the `owl:sameAs` identity predicate and its semantics. Section 5 gives an overview of existing strategies and services for managing identity in the Web of Data. Section 6 covers available solutions for the (semi-)automatic detection of erroneous identity links. Section 7 concludes by reflecting on the current state of the "sameAs problem" and identifying the most pressing and promising directions for future research.

2. Identity Problem Overview

Identity is an old and thorny topic. Classically speaking, entities that are identical are considered to share the same properties. With N denoting the set of all names, and Ψ the set of all properties, this 'Indiscernibility of Identicals' (1) is attributed to Leibniz and its converse, the 'Identity of Indiscernibles' (2) states that entities that share the same properties are identical. That identity is reflexive, symmetrical and transitive also follows from Leibniz's Law.

$$a = b \rightarrow (\forall_{\psi \in \Psi})(\psi(a) = \psi(b)) \tag{1}$$

$$(\forall_{\psi \in \Psi})(\psi(a) = \psi(b)) \rightarrow a = b \tag{2}$$

This identity relation induces a partitioning of N into a collection of non-empty and mutually disjoint *equivalence classes* $N_k \subseteq N$. From the premises $\psi(a)$, and $a, b \in N_k$, it follows that $\psi(b)$ is also the case. In fact, this deduction is central to the Web of Data as it allows complementary descriptions of the same resource to be maintained locally, yet interchanged globally, merely by interlinking the names that are used in those respective descriptions. However, there are also problems with it, and – consequently – criticisms have been levelled against it. These problems are not new, neither specific to the Web of Data, as they are present in all KR systems [10, 11]. However, the problems are specifically pressing in the Web of Data due to its unprecedented size, the heterogeneity of its content and users, and the absence of a central naming authority. This section briefly presents some of the well-known issues with this notion of identity.

2.1. Philosophical Problems

From a philosophical point of view, we present the two major issues with this notion of identity. Firstly, identity over time poses problems, as a ship[2] may still be considered the same ship even though some, or even all, of its original components (i.e. properties) have been replaced by new ones [12]. In addition, identity is context-dependent [13], allowing two medicines, having the same chemical structure, to be considered the same in a medical context, but to be considered different in other contexts (e.g. because they are produced by different companies). These issues in the classical identity definition have led to various philosophical theories, such as the distinction between accidental properties (traits that could be taken away from an object without making it a different thing), and essential properties (core elements needed for a thing to be the thing that it is) [14].

2.2. Practical Problems

Given that this problematic notion of identity is also standardised as part of the Web Ontology Language (OWL), it is normal to encounter these issues in Web applications. In fact, and due to the Open World Assumption and the continuous increase of Ψ, identity statements in the Web of Data are even more controversial. Firstly, unless two things are explicitly said to be different (e.g. using `owl:differentFrom`), the absence of an identity statement between them does not mean that they are not identical. Compared to the 558M `owl:sameAs` present in a 2015 crawl of the Web of Data [15], this type of statement is barely present in the Web of Data, with only 3.6K `owl:differentFrom` statements existing at that time in the same dataset. In addition, most `owl:sameAs` links are generated by heuristic entity resolution techniques, that employ practical strategies which are not guaranteed to be accurate. For instance, the precision of such tools ranged between 79% and 92% in the 2019 Ontology Alignment Evaluation Initiative (OAEI- SPIMBENCH track)[3]. For instance, an algorithm matching books based on the similarity of their titles and authors is not always accurate, as two different editions of the same book can also share both these traits without being `owl:sameAs` (e.g. because they do not share the same number of pages). Finally, studies have shown that modellers have different opinions about whether two objects are the same or not. For instance in [4], three KR experts were asked to judge 250 `owl:sameAs` links collected from the Web. The evaluation shows high disagreements, with one judge confirming the correctness of only 73 `owl:sameAs` statements, whilst the two other experts judging up to 132 and 181 links as true. While in some cases this may be due to differences in modelling competence, there is also the problem that two modellers may consider different parts of the same knowledge graph within different contexts.

3. Identity Analysis Approaches

The special status of `owl:sameAs` links has motivated several studies into investigating the use of these links in the Web of Data. This section presents these studies that analyses the use of `owl:sameAs` on different aspects, either by analysing its use at the aggregated level of datasets (section 3.1), studying the structure of `owl:sameAs` networks (section 3.2), or determining the quality of these existing links (section 3.3).

3.1. Dataset/Namespace Interlinking Analyses

Some studies have focused on the use of identity at the aggregated level of datasets, in order to better understand the common interests between different Linked Data publishers. In such studies, graph nodes represent the datasets, and weighted edges represent the number of `owl:sameAs` linking the dataset resources. For grouping the retrieved resources into datasets, these studies assume that all data originating from one pay-level domain (PLD) belongs to a single dataset. In an early study, the authors of [16] extracted 8.7M `owl:sameAs` triples from the

[2] Reference to the ship of Theseus or Theseus's paradox

[3] https://project-hobbit.eu/challenges/om2019/

2010 Billion Triple Challenge dataset[4]. By visualizing the largest connected component, this study shows that densely connected clusters usually represent datasets that cover similar topics (e.g. a cluster of datasets that publish data related to scientific publications, and a cluster of bioinformatics datasets). A later analysis [17] crawled 1,014 datasets containing 8M terms. The entire graph of datasets was found to consist of 9 weakly connected components with the largest one containing 297 datasets. This study shows that dbpedia.org has the largest in-degree (89 datasets asserting owl:sameAs links to DBpedia entities), and that bibsonomy.org has the largest out-degree (Bibsonomy entities are linked to 91 different datasets). The authors have also analysed the use of other linking predicates, within different topics (e.g. life sciences, geography, publications). This study shows that owl:sameAs is the most used predicate for linking within most topics, followed by rdfs:seeAlso for life sciences datasets and foaf:knows for social networking datasets. Finally, a recent study [18] have analysed 558.9M distinct owl:sameAs triples collected from a 2015 crawl of the LOD Cloud [15]. The resulting graph[5] contains 2,618 datasets, connected through 10,791 edges, and consists of 142 connected components. This study shows that there are several high-centrality nodes that act as domain-specific naming authorities/hubs in the LOD Cloud, such as geonames.org for interlinking geographic datasets, and bio2rdf.org for interlinking biochemistry datasets. This study concludes by showing that the majority of datasets have incoming links, whilst far fewer datasets have outgoing links, indicating that a relatively small number of datasets is linking to a relatively large number of them.

3.2. *Identity Graph's Structure Analyses*

Other studies have focused on analysing the graph structure of the owl:sameAs networks, where edges represent owl:sameAs and nodes represent the subjects and objects occurring in owl:sameAs triples. In a 2010 analysis [16], the transitive closure of 8.7M owl:sameAs triples have resulted in a graph of 2.9M connected components (i.e. equivalence classes). Most of these classes are small (average size of 2.4 terms), with only 41 classes with hundreds of terms, and only two classes with thousands of terms. This study also shows that owl:sameAs networks have mostly a star-like structure consisting of single central resource connected to a number of peripheral resources. In 2011, the authors of [19] extracted 3.7M distinct owl:sameAs from a corpus of 947M distinct RDF triples, crawled from 3.9M RDF/XML web-documents in 2010. After transitive closure, the data formed 2.16M equivalence classes (average size of 2.65 terms). The largest equivalence class contains 8,481 terms, with 74% of the equivalence classes containing only two terms. In a later analysis based on the 2011 Billion Triple Challenge dataset, the authors of [20] observed that the number of owl:sameAs statements per term approximates a power-law distribution with coefficient -2.528. However, in a more recent analysis of 558.9M distinct owl:sameAs statements linking 179.7M terms, the authors of [18] find that although most terms do appear in a small number of owl:sameAs statements, this distribution does not display a power-law distribution. Also in this study, the authors of [18] calculate the transitive closure of this collection of owl:sameAs statements, resulting in 48.9M non-singleton equivalence classes. With 64% of these equivalence classes containing only two terms, the size distribution of these equivalence classes fits a power law with exponent 3.3 ±0.04. On average, an equivalence class contains 3.7 terms, with the largest one containing 177,794 terms. This study also shows that the materialization of this closure would consist of 35,201,120,188 triples, and shows that only 130,673,158 owl:sameAs are necessary for obtaining the same closure (i.e. kernel). Hence suggesting that 76.6% of the existing identity statements are redundant. Finally, and in the same year, the authors of [21] studied the LOD connectivity by relying on 302 datasets collected from existing data dumps [17], datahub.io, linklion.org, and subsets of DBpedia, Wikidata, Yago, and Freebase. Based on the 44M collected owl:sameAs linking 65.3M terms, the transitive closure results in 24M equivalence classes (average size of 2.7 terms per class). Similarly to [18], the size distribution of these equivalence classes shows a power law distribution, with around 70% of the classes containing only two terms.

[4] Dataset crawled during March/April 2010 based on datasets provided by Falcon-S, Sindice, Swoogle, SWSE, and Watson using the MultiCrawler/SWSE framework

[5] Available at https://www.sameas.cc/explicit/img.svg

3.3. Quality Analysis

Finally, other type of analyses have focused on the quality of existing `owl:sameAs` links in the Web of Data. In such studies, Semantic Web experts were asked to manually judge if two IRIs, claimed to be the same, actually refer to the same real-world entity, whilst carefully considering the difference between non-information resources and information resources. This type of study was firstly conducted by [22] in 2008, in which the authors assessed the quality of authors linkage with DBpedia in the 2006 DBLP dataset. By looking at the 49 most common author names, the study shows that 92% of these authors have incorrect publications affiliated to them, due to erroneous `owl:sameAs` assertions. In 2010, the authors of [4] manually evaluated a sample of 250 `owl:sameAs` statements from a collection of 58.6M `owl:sameAs` links. This study shows that around 21% of the `owl:sameAs` assertions are incorrect, and should be replaced by a similarity or 'related to' relationships. In a follow up study [23], the authors have showed that `owl:sameAs` links resulted from inference are more likely to be erroneous than randomly chosen explicitly asserted ones. In another `owl:sameAs` quality analysis, the authors of [2] manually evaluated 1K pairs occurring in the same equivalence classes, following the transitive closure of 3.7M distinct `owl:sameAs` triples. This evaluation shows that 2.8% of the pairs are different, and should not belong to the same equivalence class. Finally, a recent analysis by [5] shows that the transitive closure of 558.9M `owl:sameAs` links result in a number of large equivalence classes, that are potentially erroneous. For instance, the largest equivalence class contains 177,794 IRIs, which in theory represents thousands of different names referring to the same real-world entity, but in practice refer to a number of different countries, cities, persons, products and activities (e.g. Bolivia, Dublin, Albert Einstein, and Basketball). In addition, based on the manual evaluation of 300 `owl:sameAs` links, while relying on the distribution of a computed error degree and the number of symmetrical `owl:sameAs`, the author of [3] estimate that around 4% of the existing `owl:sameAs` triples are erroneous.

Discussion
These different and complementary studies have investigated several aspects of the use of identity in the Web of Data. Firstly, they show that not all datasets are transitively linked in the LOD Cloud by `owl:sameAs` assertions [17, 18], with some connected components consisting of clusters of densely connected datasets that cover similar topics [16]. In addition, these studies show that `owl:sameAs` networks have a particular structure, often consisting of central IRIs connected to other peripheral ones [16, 18]. Furthermore, studies that compute the `owl:sameAs` transitive closure show that, on average, each real-world entity is represented by around three IRIs in the LOD Cloud [16, 18, 19, 21]. Finally, regarding the quality of the existing interlinks, these studies have confirmed the presence of a number of incorrect identity links in the Web of Data, with [2] and [24] estimating the number of erroneous links to 2.8% and 4% respectively, whilst [4] evaluation suggests that around one out of five `owl:sameAs` links in the Web of Data is erroneous.

In comparison to the size of the Web of Data, the conducted analyses regarding the identity links' quality are still not representative enough. With the current quality estimation being based on a maximum of 1K identical pairs, there is an obvious need for more community-based or crowd-sourcing initiatives in order to evaluate the quality of Linked Data's most essential ingredient: *identity links*. Ideally, these initiatives can be backed with (semi-) automatic approaches, that can help flagging potentially erroneous identity links. These approaches are presented and discussed in Section 6. Besides the uncertainty over the identity links' quality, there is also a technical burden preventing analyses on a larger scale than the ones presented in this section. This problem have motivated several initiatives to harvest the Web, and provide efficient access to these identity links with their transitive closure. We present and compare these approaches in Section 5. Finally, while most approaches have focused on analysing the use of `owl:sameAs` in the Web and fixing the quality of existing links, there is several approaches that introduced alternatives to the standard Semantic Web identity predicate. In the next section, we present these `owl:sameAs` alternatives, and discuss their advantages and drawbacks.

4. Alternative Identity Links

Given the philosophical and practical problems of `owl:sameAs` presented in Section 2, a number of vocabularies and approaches have acknowledged the excessive use of `owl:sameAs` and provided alternative similarity and identity links. This section presents the most deployed alternatives and gives an overview of their usage in Table 1.

4.1. Weak-Identity and Similarity Predicates

rdfs:seeAlso. This property is not used to denote any identity relation, but is used to indicate a resource that might provide additional information about the subject resource. This relationship was heavily used in linking Friend of a Friend (FOAF) data alongside the property foaf:knows, prior to the rise of owl:sameAs [25]. Despite not having well-defined semantics, this property could still be useful in linking closely related entities and datasets.

SKOS predicates. The Simple Knowledge Organisation System (SKOS) [26] is a common data model for sharing and linking knowledge organization systems via the Semantic Web. SKOS introduces three mapping properties that correspond to different types of owl:sameAs usage. Firstly, skos:relatedMatch is used to state an associative mapping link between two concepts. The predicate skos:closeMatch indicates that "two concepts are sufficiently similar that they can be used interchangeably in some applications". Finally, skos:exactMatch indicates "a high degree of confidence that the concepts can be used interchangeably across a wide range of applications". Whilst the misuse of these mapping properties can have less implications than the misuse of owl:sameAs due to their less strict semantics, their benefits for linking names is also limited due to the lack of well-defined contexts of use. For instance, skos:relatedMatch is highly ambiguous and could probably relate most the concepts of the Semantic Web (since everything is related to everything in some way). In addition, the applications (i.e. the contexts) where the concepts related by skos:closeMatch or skos:exactMatch can interchange are not defined, and are eventually subjective.

umbel:isLike. The UMBEL[6] vocabulary introduces predicates such as umbel:isLike for "asserting an associative link between similar individuals who may or may not be identical, but are believed to be so".

vocab:similarTo. The Vocab.org[7] vocabulary introduces the vocab:similarTo predicate to be "used when having two things that are not owl:sameAs but are similar to a certain extent".

lvont:strictlySameAs. For expressing near-identity, the Lexvo.org[8] vocabulary introduces the similarity predicates lvont:nearlySameAs and lvont:somewhatSameAs. According to [27], "the definitions of these predicates were intentionally left vague, simply because similarity is a very vague notion". In addition, Lexvo.org introduces lvont:strictlySameAs, a predicate which is declared formally equivalent to owl:sameAs, but just introduced for the purpose of distinguishing strict identity use from the erroneous use of the latter.

wdt:P2888. In Wikidata[9], the exact match predicate (P2888) is deployed for linking concepts, and is declared as equivalent to skos:exactMatch.

schema:sameAs. The schema.org vocabulary[10] includes the schema:sameAs property. However despite their name similarities, the semantics of this property is substantially different from that of owl:sameAs, as it states that two terms "are two pages with the same primary topic" and does not express equality.

Similarity Ontology. Finally, in order to express different types of identity relations, the authors of [4] propose the Similarity Ontology (SO) in which they hierarchically represent 13 different similarity and identity predicates. This ontology includes owl:sameAs, rdfs:seeAlso, and the three previously described SKOS predicates. For formally defining their semantics, the authors have characterised the remaining eight newly introduced predicates by reflexivity, transitivity and symmetry properties. The most specific predicate in this ontology is owl:sameAs, and the most general ones are so:claimsRelated and so:claimsSimilar. The predicates prefixed with the word claims express a subjective identity or similarity relation in which their validity depends on the (contextual) interpretation of the user. The most specific newly-introduced predicate is so:identical. This predicate follows the owl:sameAs definition, in the sense

[6] http://umbel.org

[7] http://vocab.org

[8] http://lexvo.org

[9] http://wikidata.org

[10] http://schema.org

that two IRIs linked by this predicate do refer to the same real world entity. However, and contrary to `owl:sameAs`, this predicate is referentially opaque and does not follow Leibniz's law. Meaning that properties ascribed to one IRI are not necessarily appropriate for the other, and can not be substituted. As an example of referential opacity, the authors state the case of social inappropriateness in using certain names, referring to the same real world entity, in different contexts. However, and despite proposing several alternative semantics for the strict identity relationship, this approach does not tackle the problem of how a context, where a certain identity link only holds, can be explicitly represented. Therefore, no indications on which properties ascribed to one IRI, will be also inferred to its identical or similar IRI.

4.2. Contextual Identity

The standardised semantics of `owl:sameAs` can be thought of as instigating an implicit context that is characterised by all (possible) properties to have the same values for the linked names. Weaker types of identity can be expressed by considering a subset of properties with respect to which two resources can be considered the same. At the moment, the way of encoding contexts on the Web is largely ad hoc, as contexts are often embedded in application programs, or implied by community agreement. The issue of deploying contexts in KR systems has been extensively studied in AI [28]. In the Web of Data, explicit representation of context has been a topic of discussion since its early days [29], where the variety and volume of the web poses a new set of challenges than the ones encountered in previous AI systems. This section presents approaches focusing on the specific issue of representing contextual identity in the Web.

Firstly in [30], a context Π is defined as a subset of all properties Ψ which are necessary and sufficient to determine indiscernibility and hence identity:

$$a =_{\Pi} b \rightarrow (\forall_{\pi \in \Pi})(\pi(a) = \pi(b)) \tag{3}$$

$$(\forall_{\pi \in \Pi})(\pi(a) = \pi(b)) \rightarrow a =_{\Pi} b \tag{4}$$

Looking back to the example in Section 2.1, two medicines with the same chemical structure, but produced by different companies, are identical in the context where the property π_i specifying the medicine's commercial supplier is discarded (i.e. $\pi_i \notin \Pi$).

In [31], this notion of contextual identity is encoded in RDF, and the definition of a context is extended to a sub-graph of the domain ontology called a *global context*. Specifically, a global context $\mathcal{G}$ is composed of a subset of classes $C_{\mathcal{G}}$ and properties $P_{\mathcal{G}}$ of an ontology $\mathcal{O}$, and a set of axioms which are limited to constraints on property domains and ranges. These axioms allow the parameterization of the identity criteria with respect to each class of the ontology. For instance, allowing to express that two medicines are considered identical if they have the same quantity of elements of type c_1, whilst disregarding the quantity of its other elements. The identity relation between two class instances in a global context is based on the notion of graph isomorphism of their descriptions, where an approach is proposed for automatically detecting these global contexts.

With both these approaches unclear about the treatment of properties p that do not belong to the identity context (i.e. $p \notin \Pi$ or $p \notin P_{\mathcal{G}}$), a richer definition of context was proposed by [32]. It defines a context by two sets of properties, Γ for indiscernibility and Λ for propagation:

$$a =_{(\Gamma,\Lambda)} b \rightarrow (\forall_{\gamma \in \Gamma})(\gamma(a) = \gamma(b)) \tag{5}$$

$$(\forall_{\gamma \in \Gamma})(\gamma(a) = \gamma(b)) \rightarrow a =_{(\Gamma,\Lambda)} b \tag{6}$$

$$a =_{(\Gamma,\Lambda)} b \rightarrow (\forall_{\lambda \in \Lambda})(\lambda(a) = \lambda(b)) \tag{7}$$

Principles (5) and (6) refers to the same notion of contextual identity defined in [30], whilst (7) defines the notion of *contextualised propagation*. Note that unlike Γ, indiscernibility in Λ does not determine identity. For instance, in a scientific context, two medicines sharing the same chemical structure γ_1 is enough to consider them identical, and infer that they share the same purpose λ_1. However, two medicines with the same λ_1 do not necessarily share the same γ_1. This approach extends a previous approach by [33], mainly in the way of parametrizing the propagation context Λ, and the way these contextual identity links are encoded in RDF (on the triples level instead of the graphs level).

Discussion

In section 4.1, we have presented several alternative predicates that may replace the use of `owl:sameAs` in some situations. A big downside of most of these approaches is their lack of formal semantics. For example, `skos:exactMatch` indicates a high degree of confidence that the concepts can be used interchange-

Table 1

Overview of the distinct usage of alternative identity links, based on a 2015 crawl of the Web of Data, and Wikidata for *wdt:P2888* queried in March 2020.

Property	**# Dist. Triples**	**# Dist. Terms**
`owl:sameAs`	558,943,116	179,739,567
`rdfs:seeAlso`	169,172,965	206,881,244
`wdt:P2888`	1,203,646	2,389,810
`skos:exactMatch`	566,137	346,800
`umbel:isLike`	461,054	837,040
`skos:closeMatch`	371,011	435,048
`lvont:nearlySameAs`	3,067	5,832
`vocab:similarTo`	283	554
`lvont:somewhatSameAs`	1	2
`lvont:strictlySameAs`	0	0

ably across a wide range of information retrieval applications. Whether a degree of confidence is high (enough) is subjective, and the meaning of this relation even changes over time, because information is always evolving over time. Also, some proposed alternative properties do not denote equivalence relations, which means that they are of limited use in linking and reasoning. In addition, most of these approaches require data publishers to change their modelling practice, needing a lot of momentum in order to create new datasets, or to change existing ones in order to make use of these alternative properties. As a result, and as presented in Table 1, most of these proposals lack uptake and are only used in a handful of datasets. Interestingly, we can also observe from this Table how the stricter identity relations `owl:sameAs` and `skos:exactMatch` have different characteristics than the rest of the relations. Specifically, we observe that the number of distinct RDF terms appearing in the object or subject position of such triples, is significantly lower than the total number of this type of triples. Thus, suggesting the presence of larger equivalence classes when the transitive closure of these relations is computed, compared to the other relations.

The approaches proposed by [30–33], that come up with a new context-dependent semantics for the `owl:sameAs` relation have the benefit that it does not require existing modelling practices to be changed. However, existing Linked Data tools (e.g. reasoners, programming libraries, Linked Data browsers) have little support for contextual semantics. In fact, the exact impact of contextual identity on entailment, and its feasibility at the scale of the Web has not been sufficiently investigated yet. Finally, despite the need in theory for contextual identity, the practical use of identity assertions for the purpose of interlinking may be somewhat hampered by contextual semantics approaches. With the traditional semantics of `owl:sameAs`, linked descriptions can always be shared, but with contextual semantics such descriptions can only be shared if they are asserted in compatible contexts.

5. Identity Management Services

Identity management services share the common goal of helping users or applications to identify IRIs referring to the same real world entity, and distinguish similar labels referring to different real world entities. For instance, in order to avoid using a resource referring to the river of Niger, while intending in using one referring to the country Niger, one could benefit from such services for re-using an existing universal identifier that unambiguously refers to a specific real-world entity (e.g. the river of Niger). Such type of services have a more centralised vision for identity management in the Web of Data, in which each real-world entity is referenced by a single centralised IRI. On the other hand, one can make use of other type of 'decentralised' identity management services to find all identifiers referring to the river of Niger, and discover additional descriptions. Such identity observatories can play an important role in enabling large scale identity analysis in the Web, implementing and optimizing linked data queries in the presence of co-reference [34], and detecting erroneous identity assertions [5, 27, 35, 36]. This section presents existing identity management services and discuss their advantages and limitations.

5.1. Centralised Identity Management

In the early days of the Web, it was originally conceived that resource identifiers would fall into two classes: locators (URLs) to identify resources by their locations in the context of a particular access protocol such as HTTP or FTP, and names (URNs). The latter was supposed to be the standard for assigning location-independent, globally unique, and persistent identifiers to arbitrary subjects [37]. Each identifier has a defined namespace that is registered with the Internet Assigned Numbers Authority (IANA). For instance, 'ISBN' is a registered namespace that unambiguously identifies any edition of a text-based monographic publication that is available to the public. For instance, *urn:isbn:0451450523* is a URN that identifies the book "The Last Unicorn", using the ISBN namespace. Because of the lack of a well-defined resolution mechanism, and the organizational hurdle of requiring registration with IANA, URNs are hardly used (a total of 47K URNs in the 2015 crawl of the LOD Cloud [15], with only 73 registered[11] URN namespaces with IANA at the time of writing). Since 2005, the use of the terms URNs and URLs has been deprecated in technical standards in favour of the term Uniform Resource Identifier (URI), which encompasses both, and the term Internationalised Resource Identifier (IRI) which extends the URI character set that only supports ASCI encoding.

A more recent proposal for a centrally managed naming service was proposed by [38]. This public entity name service (ENS), named Okkam[12], intends to establish a global digital space for publishing and managing information about entities. Every entity is uniquely identified with an unambiguous universal URI known as an OKKAM ID, with the idea of encouraging people to reuse these identifiers instead of creating new ones. Each OKKAM ID is matched to a set of existing identifiers (e.g. DBpedia and Wikidata IRIs), using several data linking algorithms that are available in the public entity name service hosted at http://okkam.org. For instance, the company 'Apple' has a profile with an Okkam ID[13], which is linked to other non-centrally managed IDs (e.g. `dbpedia/resource/Apple_Inc`). For each OKKAM entity, a set of attributes are collected and stored in the service for the purpose of finding and distinguishing entities from another. However, the public entity name service is no longer maintained, with no information on the number of existing entities, links, and the covered datasets by the service.

5.2. *Identity Observatories*

In recent years, three identity observatories were introduced [18, 39, 40]. These web services allow users to find for a given IRI, the list of identifiers that belong to the same equivalence class. Whilst in [18] and [40] these equivalence classes are computed based solely on the transitive closure of `owl:sameAs` triples, the Consistent Reference Service (CRS) [39] incorporates a mix of identity and similarity relationships (such as `owl:sameAs`, `umbel:isLike`, `skos:closeMatch`, and `vocab:similarTo`). This service is based on 346M triples harvested from multiple RDF dumps and SPARQL endpoints, and hosted at http://sameas.org. Since its introduction in 2009, this large collection of triples linking over 203M IRIs, and resulting in 62.6M identity bundles, has been the basis for many subsequent approaches aiming to detect erroneous identity links (e.g. [27, 35, 36]).

In 2016, the authors of [40] introduced LODsyndesis, a co-reference service hosted at http://www.ics.forth.gr/isl/LODsyndesis. This service is based on the transitive closure of 44M `owl:sameAs` triples, in which the data is harvested from existing data dumps ([17], `datahub.io`, and `linklion.org`), and subsets of `DBpedia`, `Wikidata`, `Yago`, and `Freebase`. This closure results in 24M equivalence classes, that covers more than 65M terms.

Finally, a recent identity observatory was introduced by [18], and hosted at http://sameas.cc. This service provides access to the largest collection of `owl:sameAs` statements that has been gathered from the LOD Cloud to date. This collection of 558.9M distinct `owl:sameAs` is based on the 2015 LOD Laundromat corpus [41], and contains 179M unique IRIs. It also provides access to the largest `owl:sameAs` transitive closure, which consists of 49M equivalence classes.

Discussion

Identity management services play an important role in facilitating the understanding and re-use of IRIs. However we believe that centralised naming authorities such as OKKAM, although they might be adopted within some dedicated domains and applications, they will be of limited use in the context of the Web. As acknowledged by its authors [38], encouraging people to adopt and accept such Entity Naming Systems would be challenging, as the idea of having to go through an authority in order to use a new name somewhat goes against the philosophy of the ad-hoc, and scale-free nature of the Web, where "anybody is able to say anything about anything". In addition, such systems can only be truly successful once sufficient added value over the use of non-centrally managed identifiers is provided, specifically in providing efficient and high-quality search results, and offering high coverage of real-world entities. Finally, centralizing all names into

[11] https://www.iana.org/assignments/urn-namespaces/urn-namespaces.xhtml

[12] As a variation of Occam's razor: "entities are not to be multiplied without necessity"

[13] eid-9bc2b9fd-cb41-4401-8204-6c8933010acf

Table 2

Overview of Existing Identity Observatories

	sameas.org	*LODsyndesis*	*sameas.cc*
# Terms	**203,953,936**	65,315,931	179,739,567
# Statements	346,425,685	44,028,829	**558,943,116**
# owl:sameAs	Unknown	44,028,829	**558,943,116**
# Partitions	**62,591,808**	24,076,816	48,999,148
# Eq. Classes	Unknown	24,076,816	**48,999,148**

one system would raise many privacy and security concerns, in a time where the paradigm is shifting towards more decentralization of the Web [42].

On the other hand, identity observatories are more adopted in Linked Data applications (e.g. [5, 27, 35, 36]). However, in their current architecture and status, they face some limitations. Firstly, equivalence classes in the CRS service [39] are the result of the transitive closure of a mix of identity and similarity relationships (such as `umbel:isLike` and `skos:exactMatch`). Since this service does not keep the original predicates, a user cannot identify if two terms in the same bundle are actually the same, similar or just closely related (e.g. `skos:closeMatch`). The presence of several identity and similarity relations, with different semantics, means that the overall closure is not semantically interpretable (e.g. can not be used by a DL reasoner for inferring new facts). On the other hand, the LODsyndesis' main limitation lies in the number of covered resources, being an order of magnitude smaller than the two other identity observatories. Finally, with the `sameas.cc` service being based on data crawled from the 2015 LOD Laundromat crawl, its main limitation lies in its lack of up-to-date support. Table 2 presents an overview of these services, listing the number of RDF terms, RDF statements, `owl:sameAs`, partitions and equivalence classes covered by each identity observatory. Since LODsyndesis and sameas.cc are solely based on `owl:sameAs` statements, the number of statements is identical to the number of `owl:sameAs` statements, and each graph partition represent an equivalence class. In the next section, we show how such identity observatories are used in certain approaches to tackle a different aspect of the identity problem in the Web: the quality of identity links.

6. Detection of Erroneous Identity Links

Finally, an important aspect of managing identity in the Web of Data is the detection of incorrectly asserted identity links. In order to detect such links, different kinds of information may be exploited: RDF triples related to the linked resources, domain knowledge that is described in the ontology or that is obtained from experts, or different network metrics. In this section, we present existing approaches that detect erroneous identity links, based on three –occasionally overlapping– categories of approaches: inconsistency-based (6.2), content-based (6.3), and network-based approaches (6.4). Table 3 provides a summary of these approaches, stating their characteristics, requirements, and the data in which the experiments were conducted.

6.1. Evaluation Measures

An approach of erroneous link detection can be evaluated using the classic evaluation measures of precision, recall, and accuracy. In Table 3 we present these measures as reported in each paper. These evaluation measures can be defined for the problem of detection of erroneous links as follows:

Precision. Represents the number of links classified by the approach as incorrect, and are indeed erroneous `owl:sameAs` links (True Positives), over the total number of links classified as incorrect by the approach (True Positives + False Positives).

$$Precision = \frac{TP}{TP + FP}$$

Recall. Represents the number of links classified by the approach as incorrect, and are indeed erroneous `owl:sameAs` links (True Positives), over the total number of erroneous `owl:sameAs` links available in the dataset (True Positives + False Negatives).

$$Recall = \frac{TP}{TP + FN}$$

Accuracy. Represents the number of links classified by the approach as incorrect, and are indeed erroneous `owl:sameAs` links (True Positives), and the number of validated and actually correct `owl:sameAs` links (True Negatives), over the total number of `owl:sameAs` links classified as incorrect by the approach (True Positives + False Positives), and the total number of `owl:sameAs` links validated as correct by the approach (True Negatives + False Negatives).

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

6.2. Inconsistency-based Detection Approaches

Inconsistency-based approaches hypothesise that `owl:sameAs` links that lead to logical inconsistencies have higher chances of erroneousness than logically consistent `owl:sameAs`.

6.2.1. Conflicting `owl:sameAs` and `owl:differentFrom`

The first approach for detecting erroneous identity assertions in the Web of Data was introduced by [43], who presented idMesh: a probabilistic and decentralised framework for entity disambiguation. This approach hypothesises that `owl:sameAs` and `owl:differentFrom` links published by trusted sources, are more likely to be correct than links published by untrustworthy ones. For initialising the sources' trust values, the approach relies on a reputation-based trust mechanisms from P2P networks, on online communities trust metrics, or on the used domains (e.g. closed domains such as https://www.vu.nl get higher trust values). In case no information is available, a default 0.5 value is initialised for the source. The approach detects conflicting `owl:sameAs` and `owl:differentFrom` statements based on a graph-based constraint satisfaction problem that exploits the `owl:sameAs` symmetry and transitivity. They resolve the detected conflicts based on the iteratively refined trustworthiness of the sources declaring the statements (i.e. creating an autocatalytic process where constraint-satisfaction helps discovering untrustworthy sources, and where trust management delivers in return more reasonable prior values for the links). The approach shows high accuracy (75 to 90%) in discovering the equivalence and non-equivalence relations between entities even when 90% of the sources are actually spammers feeding erroneous information. However, this type of approach requires the presence of a large number of `owl:differentFrom` statements, which is not the case in the LOD Cloud (see Section 2.2). In addition, scalability evaluation, only conducted on synthetic data, demonstrate a maximum scale involving 8,000 entities and 24,000 links, over 400 machines, focusing solely on network traffic and message exchange as opposed to time. The precision and recall are not reported.

6.2.2. Ontology Axioms Violation

In [2], the authors introduced a scalable entity disambiguation approach based on detecting inconsistencies in equivalence classes. This approach detects inconsistent equivalence classes, by exploiting ten OWL 2 RL/RDF rules expressing the semantics of axioms such as *differentFrom, AsymmetricProperty, complementOf*. When resources causing inconsistencies are detected, they are separated into different seed equivalence classes. Then the approach assigns the remaining resources into one of the seed equivalence classes based on their minimum distance in the non-transitive equivalence class, or using in a case of tie, a concurrence score that is based on the pairs' shared inter- and intra- links. The authors have evaluated their approach on a set of 3.7M unique `owl:sameAs` triples derived from a corpus of 947M unique triples, crawled from 3.9M RDF/XML web-documents in 2010. From the resulting 2.8M equivalence classes, the approach detects only three types of inconsistencies in a total of 280 equivalence classes: 185 inconsistencies through disjoint classes, 94 through distinct literal values for inverse-functional properties, and one through *owl:differentFrom* assertions. On average, repairing an equivalence class requires partitioning it into 3.23 consistent partitions. After manually evaluating 503 pairs randomly chosen from the 280 inconsistent classes, the results show that 85% of the pairs that were separated from the same equivalence class are indeed erroneous (i.e. precision), leading to the separation of 40% of the pairs evaluated as erroneous by the judges (i.e. recall). This result shows that consistency does not imply correctness, with 60% of the pairs evaluated as different still belong to the same, now consistent, equivalence classes. Hence suggesting that the recall could be much lower than 40%, as the approach is not capable of detecting different pairs from the other 2.8M consistent equivalence classes. The total runtime of this approach is around 2 hours.

The authors of [44] introduced another inconsistency-based approach to invalidate identity statements. This approach firstly builds a contextual graph of a specified depth that describes each of the involved resources in a certain identity link. This contextual graph considers only the subpart of RDF descriptions that can be involved in conflicting statements: class disjointness, (inverse) functional properties and local complete properties. When the two concerned resources belong to heterogeneous sources, the approach requires the mapping of their properties. After building the contextual graphs, the Unit-resolution inference rule is applied until saturation to detect inconsistencies within these graphs. The evaluation of the approach was not based on a sample of existing `owl:sameAs` links in the Web. Instead, the authors opted for three `owl:sameAs` datasets produced by three different

Table 3

Overview of erroneous identity links detection approaches, stating their type, requirements, the dataset on which the experiments were conducted, and the reported results.

Approach	Type of Approach	Requirements	Evaluated Data	Results
[43]	Inconsistency-based	- Source trustworthiness - Presence of owl:differentFrom	Synthetic graph of 8K entities and 24K links	75% to 90% accuracy
[2]	Inconsistency-based	Ontology axioms	3.77M `owl:sameAs` from a 2010 crawl of 3.9M Web documents	85% precision, 40% recall (only 280 inconsistent classes out of 2.8M)
[44]	Inconsistency-based and Content-based	- Ontology axioms - Ontology mappings	344 `owl:sameAs` produced by 3 different linking tools (OAEI 2010)	37% to 88% precision, 75% to 100% recall (depending on the dataset)
[27]	Inconsistency-based	UNA	BTC2011: 3.4M `owl:sameAs` and sameAs.org: 22.4M `owl:sameAs`	no precision or recall evaluation
[36]	Inconsistency-based	UNA	LinkLion: 19.2M `owl:sameAs`	no precision or recall evaluation
[45]	Content-based (outlier detection)	-	Peel-DBpedia: 2K `owl:sameAs` DBTropes-DBpedia: 4.2K `owl:sameAs`	- 58% to 80% AUC - 50% F1-measure
[46]	Content-based (crowdsourcing)	Necessary descriptions for each resource	DBpedia-Freebase: 95 `owl:sameAs`	- 94% accuracy - 0% recall (higher recall for other interlinks)
[35]	Content-based (natural language analysis)	Textual description for each resource	sameas.org: 411 `owl:sameAs` (from 7K collected ones before cleansing)	93% precision 75% recall
[47]	Network Metrics (local network)	-	SILK framework: 100 `owl:sameAs`	49% precision 68% recall
[5]	Network Metrics (identity network)	-	558.9M `owl:sameAs` from a 2015 crawl of the Web of Data	93% recall, 40% to 73% precision (depending on the eq. class size)

linking tools in the context of the 2010 Ontology Alignment Evaluation Initiative (OAEI)[14], with a total of 344 links. The results show low precision in two datasets (37% and 42.3%) and high precision in the third one (88%), with a recall varying between 75 and 100% depending on the dataset. Finally, the authors show that when applied after an entity linking tool, this invalidation approach can increase the tool's precision (from 3 to 25 percentage points). However, this approach requires the presence of expert knowledge, ontology axioms, and available ontology alignment. Finally, being tested solely on a set of 344 `owl:sameAs` links, the scalability of this approach is yet to be evaluated.

6.2.3. *Unique Name Assumption Violation*

This category of approaches hypothesises that individual datasets preserve the Unique Name Assumption (UNA), and that violations of the UNA are indicative of erroneous identity links [27, 36]. The UNA indicates that two terms, with distinct IRIs in the same dataset, do not refer to the same real world entity.

The approach proposed by [27], creates undirected graphs from existing `owl:sameAs` links before applying a linear program relaxation algorithm. This algorithm aims at deleting the minimal number of edges in order to ensure that the unique name constraint is no longer violated, and is applied separately on each connected component. For the evaluation of the approach, they have firstly considered the 2011 Billion Triple Challenge dataset containing 3.4M `owl:sameAs` links, that resulted into 1.3M equivalence classes (i.e. connected components). Then a 2011 dump of the sameas.org dataset that contains 22.4M `owl:sameAs`, resulting in 11.8M equivalence classes. Finally, a third graph consisting of the combination of both data collections, containing 34.4M `owl:sameAs`, that have resulted in 12.7M equivalence classes. On the latter graph, the approach have detected 519K distinct pairs that occur in the same equivalence class, and at the same time belong to the same dataset (UNA violation). For satisfying the UNA

[14] http://oaei.ontologymatching.org/2010/

constraint, the approach removed 280K links, that represent in that context the erroneous `owl:sameAs` statements. Meaning that on average each deleted link have caused 1.85 violations in this graph, while every deleted link in the BTC2011 and sameas.org datasets have respectively caused 4.24 and 1.53 violations on average. The total runtime of the approach is not stated.

A recent approach proposed by [36] generates the equivalence classes based on an algorithm called *Union Find*. After generating the equivalence classes, and akin to [27], this approach detects the IRIs which share the same equivalence class and at the same time share the same dataset. However, instead of deleting triples to ensure the non-violation of the unique name constraint, this approach ranks the erroneous candidates based on the number of detected resources with errors. It was applied to check which link discovery framework from the LinkLion linkset repository, containing 19.2M `owl:sameAs` links, has a better score. The results show that at least 13% of the `owl:sameAs` links are “erroneous”, with sameas.org having the worst consistency, considering that the UNA is indeed respected in the LOD. The approach is scalable, with a total runtime of around 4 minutes.

The precision, recall and accuracy of both approaches have not been evaluated. Interestingly, [27] claims that most of the unique name assumption violations stem from incorrect identity links, not from inadvertent duplicates (e.g. very few DBpedia IRIs with different names exist that describe exactly the same real world entity). Whilst in the manual analysis of a random sample of 100 errors, the authors of [36] show that 90% of the errors stem from duplications within the dataset, instead of referring to two different real world entities. These contradicting interpretations leave many uncertainties on the effectiveness of the UNA assumption, within each dataset, for the task of detecting erroneous links.

6.3. Content-based Approaches

This category of approaches exploit the descriptions associated to each resource for evaluating the correctness of an identity link.

In [46], the authors looked into the use of crowdsourcing as a mean to handle data quality problems in DBpedia. The paper focuses on three categories of quality issues: (i) objects incorrectly or incompletely extracted, (ii) data types incorrectly extracted, and most importantly for the topic of this survey (iii) interlinking. The adopted methodology consists of firstly involving domain experts for finding and classifying incorrect triples, then verifying these classifications using the Amazon Mechanical Turk (MTurk). The experts flagged as incorrect a total of 1.5K triples, whilst stating each type of detected error. These triples were also evaluated by the paper’s authors as a way to create a gold standard, and were sent to the MTurk crowd for verification. Surprisingly, and according to the gold standard, Linked Data experts showed a 15% precision in evaluating interlinks. Specifically, the experts have incorrectly flagged all `owl:sameAs` statements (95 `owl:sameAs` in total, all correct, indicating a 0% precision). Checking the types of error signalled by the experts in this evaluation[15], one can see that most of these `owl:sameAs` links were signalled by the same expert, stating the same error type as “Links to Freebase". The MTurk workers have correctly judged 62% of the interlinking statements using a ‘first answer’ approach, and 94% of them using a ‘majority voting’ approach. These results show that MTurk workers can be reliable for evaluating interlinks, specially when a ‘majority voting’ approach is deployed. In addition, this work shows that finding and classifying incorrect interlinks is more complex than other types of errors. However, with the whole process taking around 25 days[16], this adapted crowdsourcing methodology is almost impossible to be applied at the scale of the LOD Cloud.

In [45], the author presented a multi-dimensional and scalable outlier detection approach for finding erroneous identity links. This work hypothesises that identity links follow certain patterns, therefore links that violate those patterns are erroneous. This approach represents each identity link as a feature vector using (i) direct types, (ii) all ingoing and outgoing properties, or (iii) a combination of both. For detecting outliers, six different methods were tested (e.g. k-NN global anomaly score, one-class support vector machines), using different parameters (10 different runs in total). Each method assigns a score to each `owl:sameAs` indicating the likeliness of being an outlier. These methods were tested on two link sets: Peel Session-DBpedia (2,087 links) and DBTropes-DBpedia (4,229 links). The experiments show much better results on

[15] https://docs.google.com/spreadsheets/d/15u3NjomX3nYF6OuMNU3w76yd5IWAcRcsTlHbCBLw6l8/edit#gid=0

[16] three predefined weeks for the contest and four days for the MTurk workers

the first dataset in terms of AUC[17], and show that using only the type features works best. The maximum F1-measure obtained is 54%, which the author states that is mainly due to flagging up to 3/4 of all links as outliers (high recall value). The precision and recall are not reported. The approach is fast in most cases, depending on which outlier detection method is applied, with a runtime varying between seconds to 15 minutes.

In [35], the authors proposed the SCID approach, that hypothesises that an `owl:sameAs` link between two resources that do not have similar textual descriptions is erroneous. This approach firstly calculates a similarity score between the IRIs involved in a given `owl:sameAs` link using the textual description associated to them (e.g., through the `rdfs:comment` property). For calculating the similarity score, the approach relies on the position and the relevance of each resource with respect to the associated DBpedia categories and then employs this score to determine whether the identity link is valid or needs to be flagged for removal. The approach was tested on 411 `owl:sameAs` links, resulting from a data cleansing of an original 7,690 link dataset extracted from sameas.org. The experimental results show that this approach can correctly flag questionable identity assertions, attaining a precision as high as 100% with a 56% recall when the threshold is set at 0.2. For a reasonable precision versus recall trade-off, the authors suggest a 0.5 or 0.6 threshold where the precision is between 86% and 93% and the recall between 75% and 79%. However, this approach requires the presence of textual descriptions for both resources, which explains the high number of discarded links from the original dataset. The evaluation was restricted on the qualitative part, without any mention on the method's scalability or the total runtime of the experiments.

6.4. Network-based Approaches

Some approaches have looked into the use of network metrics for evaluating the quality of `owl:sameAs` links.

The authors of [47] introduced LINK-QA: an extensible framework for performing quality assessment on the Web of Data. This approach, hypothesises that the quality of an `owl:sameAs` link can be determined by its impact on the network structure. This impact is measured using three classic network metrics (clustering coefficient, betweenness centrality, and degree) and two Linked Data-specific ones (`owl:sameAs` chains, and description richness). For instance, the measure of betweenness centrality is based on the idea that networks dominated by highly central nodes are more prone to critical failure in case those central nodes cease to operate or are renamed. Hence, a link's quality is calculated with respect to its impact in reducing the overall discrepancy among the centrality values of the nodes. The two Linked Data specific measures hypothesise that the quality of an `owl:sameAs` statement is measured based on its impact in closing an open `owl:sameAs` chain, and its contribution in adding complementary descriptions to the identity statement subject from the target resource. The experiments were conducted on 100 known good and bad quality links created using the Silk mapping tool. These experiments show that the classic network metrics are insufficient for assessing the quality of a link, while the impact of closing an open `owl:sameAs` chain proved more successful in distinguishing between correct and incorrect links. According to the authors, the demonstrated result of 50% precision and 68% recall is mainly due the small network sample that was chosen for the experiments. The authors claim that the approach is scalable and can be distributed, but do not state the runtime of the experiments.

Finally, we investigated in 2018 [5] the use of the community structure of the `owl:sameAs` network for detecting erroneous `owl:sameAs`. This approach hypothesises that a group of nodes that are heavily connected by `owl:sameAs` links (i.e. a community) have more chances of referring to the same real world entity, than sparsely connected ones. This approach starts by partitioning the `owl:sameAs` network into different connected components. Then, it detects the community structure of each connected component using the *Louvain* community detection algorithm [48]. Finally, it calculates an error degree for each `owl:sameAs` link. This error degree is based on the density of the community in which an `owl:sameAs` occurs in (or communities when an `owl:sameAs` is linking two terms from two different communities), and the weight of the `owl:sameAs` (i.e. reciprocally asserted `owl:sameAs` have lower error degree, hence a higher chance of correctness). The experiments were conducted on the previously described sameas.cc dataset, containing 558.9M `owl:sameAs` statements. The manual evaluation of around 300 `owl:sameAs` links shows that the precision of the approach depends on the size of the equivalence class, varying between 40%

[17] area under the ROC curve: the probability of wrong links to get lower scores than correct ones

and 73%. The recall of this approach is evaluated by injecting 780 erroneous `owl:sameAs`, suggesting a recall of 93%. The total runtime is 11 hours.

Discussion

It has now been broadly acknowledged that erroneous identity links are present in the Linked Open Data, and that additional efforts are needed in order to detect them. This section discusses the advantages and drawbacks of the presented approaches, according to the three following criteria:

Efficiency. An efficient approach is able to detect a large number of erroneous identity statements (i.e. high recall), without incorrectly classifying correct identity ones as erroneous (i.e. high precision).

Transparency. It is necessary to have approaches offering transparency to the community, by making their tools, experimental data, and their results publicly accessible. This will allow users to directly benefit from such approaches by discarding the links that were evaluated as incorrect during this approach, or only consider the ones that were validated as correct. In addition, and since probably no approach would single handedly resolve the identity links problem in the LOD, it is important to provide transparency for allowing other approaches to compare, and hopefully improve, their results. Table 4 presents the resources that were made available by each approach.

Feasibility on the LOD. According to the fourth Linked Data principle[18], the importance of identity links is its ability to interlink resources in the context of the Web of Data, and allow applications to use these links and discover new things. Hence, an important criteria is the feasibility of an approach in the context of the Linked Open Data, where approaches are expected to scale to hundreds of millions of triples, and where certain assumptions on the data can not be presumed.

Around half of the here presented approaches have looked into inconsistency detection as a mean to detect erroneous identity links. Some of these approaches are based on axioms that can be declared in the ontology, mappings that can be detected between schemas, or conflicting statements (i.e. `owl:sameAs` with `owl:differentFrom`). However, the evaluation conducted in [2] suggests that consistency does not necessarily imply correctness, showing that a large number of incorrect identity statements occur in consistent equivalence classes. In addition, these experiments show that such inconsistencies are not frequent in the LOD Cloud, with only 280 equivalence classes being inconsistent out of 2.8M classes (0.01%). This fact might have prompted other inconsistency-based approaches such as [43] and [44] to respectively conduct their experiments on synthetic data and linksets. Nevertheless, and despite the low feasibility on the LOD Cloud, these approaches have showed promising results on the respective datasets in terms of accuracy and precision, with [43] reporting an accuracy as high as 90%, [2] reporting an 85% precision, and [44] reporting an 88% precision in one linkset. However, and as presented in Table 4, these approaches offer very little transparency, as we are solely able to access the public linkset used in one experiment [44].

Other types of approaches have looked into detecting inconsistencies by presuming the unique name assumption [27, 36]. The experiments show contradicting results on whether the UNA is presumed in each dataset or not (with [27] claiming that most UNA violations stem from incorrect identity links, whilst the analysis in [36] shows that 90% of UNA violations stem from duplications). With no evaluation of the precision, recall and accuracy of both approaches, these experiments leave many uncertainties on the effectiveness of the UNA for detecting erroneous identity links.

Content-based approaches such as [46] have looked into the use of crowdsourcing for handling data quality problems in the Web, including wrong interlinks. This approach shows good efficiency in terms of precision, and offers full transparency by testing their methodology on a public dataset, and providing access to their tool, results, and gold standard. However, and as expected, crowdsourcing approaches are not scalable, requiring around 25 days for flagging a total of 1.5K DBpedia triples. On the other hand, automated content-based approaches such as [35] have showed promising results by associating the resources' textual descriptions with DBpedia categories for understanding the linked resources' meaning. Despite reporting recall numbers as high as 90%, the experiments suggest that recall is much lower in the context of the LOD Cloud, as they were able to conduct the experiments on only 411 `owl:sameAs` out of 7,690 initially picked (due to a preliminary data cleansing that mainly discards resources with no textual descriptions). In addition, and

[18] https://www.w3.org/DesignIssues/LinkedData.html

Table 4

Transparency overview of each erroneous identity links detection approach. Links are available as end notes at the end of the paper.

Approach	Dataset	Tool	Results	Gold Standard
[43]	-	-	-	-
[2]	-	-	-	Link not Working[1]
[47]	File Dumps[2]	Source Code[3]	HTML Reports[4]	-
[27]	BTC 2011[5]	-	-	-
[46]	DBpedia[6]	Source Code[7]	- Campaign Results[8] - MTurk Results[9]	Authors Evaluation[10]
[44]	PR OAEI 2010[11]	-	-	-
[45]	- Peel Sessions[12] - DBTropes[13]	Workflow[14]	-	-
[35]	-	One Function but Link not Working[15]	-	-
[36]	Link not Working[16]	Source Code[17]	-	100 Output Samples[18]
[49]	LOD crawl[19]	Source Code[20]	Box Plots[21]	-
[5]	- LOD-a-lot dataset[22] - sameAs.cc dataset[23]	Source Code[24]	556M owl:sameAs links with their error degrees[25]	300 manually evaluated links[26]

since there is no mention of the total runtime of this approach, the feasibility of this approach on millions of RDF triples (more likely billions, since they also require additional triples than owl:sameAs links) has not been demonstrated. Other content-based approaches such as [45] have showed that the resources' types can be exploited for detecting outlier identity links, with AUC as high as 80%, and an F1-measure of 50%. However, the experiments suggest low precisions, with the reported results showing that in certain cases, up to 3/4 of all links are flagged as outliers. In addition, the experiments show significant differences between the reported results in each dataset (with AUC dropping from 80% to 58% in the DPTropes dataset). Hence, indicating that such methods are highly dependant on how the data are modelled. Finally, with the approach being tested on around 6K links, its feasibility on the LOD Cloud is yet to be evaluated.

Finally, the two remaining approaches [5, 47] have looked into the use of network metrics for evaluating the quality of owl:sameAs links, without requiring assumptions on the data. The experiments in [47] on a sample of 100 links, show that classic network metrics (clustering coefficient, betweenness centrality, and degree) are not efficient for evaluating the quality of an owl:sameAs link. The Linked Data specific network metrics that are based on closing owl:sameAs chains have been proven to be slightly more effective. On the other hand, the approach proposed by [5] that is based on the community structure of each connected component of the owl:sameAs network, is the first approach that assigns an error degree to such a large collection of identity links. However, the experiments suggest that such approach can only be successfully when applied on connected components with a relatively large number of terms and links. However from the sameAs.cc dataset, we can observe that 64% of the equivalence classes in the LOD Cloud contain only two terms [18].

7. Conclusion and Discussion

This survey has presented the first overview in the ongoing process of limiting the excessive and incorrect use of identity links in the Web of Data. We now present the current situation, and set out directions for future work.

Alternative identity links lack semantics. In Section 4.1, several alternative identity and similarity predicates were presented. A big downside of these alternatives is their lack of formal semantics. For instance, in skos:exactMatch whether a degree of confidence is high (enough) is subjective, and the meaning of this relation even changes over time, because information is always evolving over time. Also, some proposed alternative properties do not denote equivalence relations, which means that they are of limited use in reasoning and linking. Another downside of these approaches is that they require data publishers to change their modelling practice. A lot of momentum is needed in order to create new knowledge graphs, or to change existing ones in order to make use of these alternative properties. As a result, most of these proposals lack uptake and are only used in a handful of datasets (see Table 1).

Contextual identity requires further investigation. In Section 4.2, different proposals for context-dependent semantics of identity were presented. These approaches have the benefit that they do not require existing modelling practices to be changed since the same property (i.e., `owl:sameAs`) can be used. An exception to this are approaches that require contexts to be modelled by hand. However, contextual semantics has not yet been widely implemented in Linked Data tools, e.g., reasoners, linked data browsers, and faces potential impediments for uptake. In fact, the exact impact of contextual identity on entailment has not been sufficiently investigated. Finally, the use of identity assertions for the purpose of interlinking may be somewhat hampered by contextual semantics approaches. With the traditional semantics of `owl:sameAs`, linked descriptions can always be shared, but with contextual semantics such descriptions can only be shared if they are asserted in compatible contexts.

Centralised naming authorities will be of limited use. Centralised naming authorities, presented in Section 5.1, play an important role in facilitating the understanding and re-use of names. However, although they might see limited uptake within some dedicated domains, centralised identity management becomes more difficult and error prone when operating at a larger scale. In addition, the idea of having to go through an authority in order to use a new name somewhat goes against the philosophy of the ad hoc nature of the Web, where "anybody is able to say anything about anything".

Identity Observatories must be used more broadly. Even though several identity observatories exist (Section 5.2), they are not commonly used in Web applications today. This is probably due to the following limitations which these services suffer from, in their current status and architecture.

Semantic Interpretability. The 'equivalence classes' in sameas.org are the result of the transitive closure of a mix of identity relations with different semantics. Since this service does not keep the original predicates, the semantics of the closure that is calculated is unclear (e.g. can not be used by a DL reasoner for inferring new facts).

Coverage. With the number of statements in LODsyndesis being an order of magnitude smaller than other observatories, this service may see limited use in certain applications.

Up-to-date support. With sameas.cc being based on a 2015 crawl of the Web, such service may see limited uptake in applications which require more recent information.

We believe that such services will see uptake over time, since they make it possible to use some of the benefits of linking to other knowledge graphs, while at the same time giving the client some control as to which knowledge graphs to link to (and which ones not to link to).

Hybrid error detection approaches are required. Finally, it has now been broadly acknowledged that erroneous identity statements are present in the Web of Data, and that additional effort is needed in order to detect them. In Section 6, we have seen that there are several promising approaches for the (semi-) automatic detection of erroneous identity links. However, all existing approaches have made some trade-off, either having less precision, having less recall, or being less scalable. Specifically, experiments in [2] showed that the Web of Data lack from ontological axioms and assertions that are strong enough for deriving inconsistencies. Hence, suggesting that axiom violation-based approaches will mainly have a lower recall. Experiments based on the UNA violation have showed contradicting results, leaving many uncertainties on the effectiveness of the UNA assumption for the task of detecting erroneous links. Content-based approaches have showed promising results in terms of precision and recall, but still requires further investigation for testing their scalability, and determining whether sufficient textual descriptions in the Web of Data are indeed available. Finally, network-based approaches have also showed promising results in terms of recall and scalability, but existing experiments shows lower precision.

Future research should focus on combining some of these existing approaches in novel ways, potentially combining some of the strengths of these various approaches into one (hybrid) approach. Such an approach should be feasible over the LOD Cloud, where scalability is not the only challenge, but also where certain assumptions on the constant changing data can not be presumed. For instance, in the LOD Cloud not all names have textual descriptions, many knowledge graphs do not include vocabulary mappings, or lack semantically rich assertions for deriving inconsistencies. In addition, future research should focus on providing more transparency for allowing other approaches to compare, and hopefully improve, their results. Table 3 shows that only three approaches provide fully reproducible results. Finally, compared to the amount of research invested in entity linking [50] and ontology matching [6], this area is clearly lacking uptake. While in some cases this may be due to various technical challenges (e.g. resulted from the absence of manually annotated benchmarks designed for this task), there is also the aspect that the number and actual effects of these erroneous statements in practice were still unknown, until recently [5].

With this overview on the current state of the "sameAs problem", we hope that this survey can lead to the emergence of more efficient approaches and systems for representing contextual identity and investigating their impact at scale, accessing explicit and implicit identity assertions in the Web, and detecting the erroneous ones.

References

[1] S. Bechhofer, F. Van Harmelen, J. Hendler, I. Horrocks, D.L. McGuinness, P.F. Patel-Schneider, L.A. Stein et al., OWL web ontology language reference, *W3C recommendation* **10**(02) (2004).

[2] A. Hogan, A. Zimmermann, J. Umbrich, A. Polleres and S. Decker, Scalable and distributed methods for entity matching, consolidation and disambiguation over linked data corpora, *Web Semantics: Science, Services and Agents on the World Wide Web* **10** (2012), 76–110.

[3] J. Raad, Identity Management in Knowledge Graphs (doctoral dissertation), University of Paris-Saclay, 2018.

[4] H. Halpin, P.J. Hayes, J.P. McCusker, D.L. McGuinness and H.S. Thompson, When owl:sameAs isn't the same: An analysis of identity in Linked Data, in: *International Semantic Web Conference*, Springer, 2010, pp. 305–320.

[5] J. Raad, W. Beek, F. van Harmelen, N. Pernelle and F. Saïs, Detecting Erroneous Identity Links on the Web Using Network Metrics, in: *International Semantic Web Conference*, Springer, 2018, pp. 391–407.

[6] A. Ferrara, A. Nikolov and F. Scharffe, Data linking for the semantic web, *Semantic Web: Ontology and Knowledge Base Enabled Tools, Services, and Applications* **169** (2013), 326.

[7] H. Halpin and V. Presutti, An ontology of resources: Solving the identity crisis, in: *European Semantic Web Conference*, Springer, 2009, pp. 521–534.

[8] H. Halpin, Sense and Reference on the Web (doctoral dissertation), University of Edinburgh, 2010.

[9] H. Halpin and A. Monnin, *Philosophical Engineering: Toward a Philosophy of the Web*, John Wiley & Sons, 2013.

[10] J. Grant and V.S. Subrahmanian, Reasoning in Inconsistent Knowledge Bases, *IEEE Trans. Knowl. Data Eng.* **7**(1) (1995), 177–189. doi:10.1109/69.368510. https://doi.org/10.1109/69.368510.

[11] N.T. Nguyen, *Advanced Methods for Inconsistent Knowledge Management (Advanced Information and Knowledge Processing)*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2007. ISBN 1846288886.

[12] D. Lewis, On the plurality of worlds, *Oxford* **14** (1986), 43.

[13] P.T. Geach, Identity, *Review of Metaphysics* **21** (1967), 3–12.

[14] S.A. Kripke, Naming and necessity, in: *Semantics of natural language*, Springer, 1972, pp. 253–355.

[15] J.D. Fernández, W. Beek, M.A. Martínez-Prieto and M. Arias, LOD-a-lot, in: *International Semantic Web Conference*, Springer, 2017, pp. 75–83.

[16] L. Ding, J. Shinavier, Z. Shangguan and D.L. McGuinness, SameAs networks and beyond: analyzing deployment status and implications of owl: sameAs in linked data, in: *International Semantic Web Conference*, Springer, 2010, pp. 145–160.

[17] M. Schmachtenberg, C. Bizer and H. Paulheim, Adoption of the linked data best practices in different topical domains, in: *International Semantic Web Conference*, Springer, 2014, pp. 245–260.

[18] W. Beek, J. Raad, J. Wielemaker and F. van Harmelen, sameAs. cc: The Closure of 500M owl: sameAs Statements, in: *Extended Semantic Web Conference*, Springer, 2018, pp. 65–80.

[19] A. Hogan, A. Harth, J. Umbrich, S. Kinsella, A. Polleres and S. Decker, Searching and browsing linked data with swse: The semantic web search engine, *Web semantics: science, services and agents on the world wide web* **9**(4) (2011), 365–401.

[20] X. Wang, T. Tiropanis and H.C. Davis, Optimising Linked Data Queries in the Presence of Co-reference, in: *European Semantic Web Conference*, Springer, 2014, pp. 442–456.

[21] M. Mountantonakis and Y. Tzitzikas, Scalable Methods for Measuring the Connectivity and Quality of Large Numbers of Linked Datasets, *Journal of Data and Information Quality (JDIQ)* **9**(3) (2018), 15.

[22] A. Jaffri, H. Glaser and I. Millard, URI Disambiguation in the Context of Linked Data, in: *WWW Workshop on Linked Data on the Web, LDOW*, 2008.

[23] H. Halpin, P.J. Hayes and H.S. Thompson, When owl: sameAs isn't the same redux: towards a theory of identity, context, and inference on the semantic web, in: *International and Interdisciplinary Conference on Modeling and Using Context*, Springer, 2015, pp. 47–60.

[24] J. Raad, W. Beek, N. Pernelle, F. Saïs and F. van Harmelen, Détection de liens d'identité erronés en utilisant la détection de communautés dans les graphes d'identité, in: *Revue des Sciences et Technologies de l'Information-Série ISI: Ingénierie des Systèmes d'Information*, 2018.

[25] L. Ding, J. Shinavier, T. Finin, D.L. McGuinness et al., owl: sameAs and Linked Data: An empirical study, in: *Proceedings of the Second Web Science Conference*, 2010.

[26] A. Miles and S. Bechhofer, SKOS Simple Knowledge Organization System Reference. W3C Recommenda-

tion 18 August 2009., 2009. http://www.w3.org/TR/2009/REC-skos-reference-20090818/.

[27] G. de Melo, Not Quite the Same: Identity Constraints for the Web of Linked Data, in: *The Twenty-Seventh AAAI Conference on Artificial Intelligence*, AAAI Press, 2013.

[28] R.V. Guha, *Contexts: a formalization and some applications*, Vol. 101, Stanford University Stanford, CA, 1991.

[29] P. Bouquet, F. Giunchiglia, F. Van Harmelen, L. Serafini and H. Stuckenschmidt, C-owl: Contextualizing ontologies, in: *International Semantic Web Conference*, Springer, 2003, pp. 164–179.

[30] W. Beek, S. Schlobach and F. van Harmelen, A Contextualised Semantics for owl: sameAs, in: *International Semantic Web Conference*, Springer, 2016, pp. 405–419.

[31] J. Raad, N. Pernelle and F. Saïs, Detection of Contextual Identity Links in a Knowledge Base, in: *Proceedings of the Knowledge Capture Conference*, ACM, 2017, p. 8.

[32] A.K. Idrissou, R. Hoekstra, F. van Harmelen, A. Khalili and P. van den Besselaar, Is my: sameAs the same as your: sameAs?: Lenticular Lenses for Context-Specific Identity, in: *International K-CAP Conference*, ACM, 2017, p. 23.

[33] C. Batchelor, C.Y. Brenninkmeijer, C. Chichester, M. Davies, D. Digles, I. Dunlop, C.T. Evelo, A. Gaulton, C. Goble, A.J. Gray et al., Scientific lenses to support multiple views over linked chemistry data, in: *International Semantic Web Conference*, Springer, 2014, pp. 98–113.

[34] K. Schlegel, F. Stegmaier, S. Bayerl, M. Granitzer and H. Kosch, Balloon fusion: SPARQL rewriting based on unified co-reference information, in: *Data Engineering Workshops (ICDEW), 2014 IEEE 30th International Conference on*, IEEE, 2014, pp. 254–259.

[35] J. Cuzzola, E. Bagheri and J. Jovanovic, Filtering inaccurate entity co-references on the linked open data, in: *International DEXA Conference*, Springer, 2015, pp. 128–143.

[36] A. Valdestilhas, T. Soru and A.-C.N. Ngomo, CEDAL: time-efficient detection of erroneous links in large-scale link repositories, in: *International Conference on Web Intelligence*, ACM, 2017, pp. 106–113.

[37] M. Mealling and R. Daniel, URI Resolution Services Necessary for URN Resolution (RFC 2483), IETF, 1999. http://www.ietf.org/rfc/rfc2483.txt.

[38] P. Bouquet, H. Stoermer and D. Giacomuzzi, OKKAM: Enabling a Web of Entities., *I3* **5** (2007), 7.

[39] H. Glaser, A. Jaffri and I. Millard, Managing Co-reference on the Semantic Web, in: *Proceedings of the WWW Workshop on Linked Data on the Web, LDOW*, 2009.

[40] M. Mountantonakis and Y. Tzitzikas, On measuring the lattice of commonalities among several linked datasets, *Proceedings of the VLDB Endowment* **9**(12) (2016), 1101–1112.

[41] W. Beek, L. Rietveld, H.R. Bazoobandi, J. Wielemaker and S. Schlobach, LOD laundromat: a uniform way of publishing other people's dirty data, in: *International Semantic Web Conference*, Springer, 2014, pp. 213–228.

[42] R. Verborgh, T. Kuhn and A. Sambra (eds), Proceedings of the Workshop on Decentralizing the Semantic Web, in: *Proceedings of the Workshop on Decentralizing the Semantic Web*, CEUR Workshop Proceedings, Aachen, 2017, ISSN 1613-0073. http://ceur-ws.org/Vol-1934/.

[43] P. CudreMauroux, P. Haghani, M. Jost, K. Aberer and H. De Meer, idMesh: graph-based disambiguation of linked data, in: *International conference WWW*, ACM, 2009, pp. 591–600.

[44] L. Papaleo, N. Pernelle, F. Saïs and C. Dumont, Logical detection of invalid SameAs statements in RDF data, in: *International Conference EKAW*, Springer, 2014, pp. 373–384.

[45] H. Paulheim, Identifying Wrong Links between Datasets by Multi-dimensional Outlier Detection., in: *WoDOOM*, 2014, pp. 27–38.

[46] M. Acosta, A. Zaveri, E. Simperl, D. Kontokostas, S. Auer and J. Lehmann, Crowdsourcing linked data quality assessment, in: *International Semantic Web Conference*, Springer, 2013, pp. 260–276.

[47] C. Guéret, P. Groth, C. Stadler and J. Lehmann, Assessing linked data mappings using network measures, in: *Extended Semantic Web Conference*, Springer, 2012, pp. 87–102.

[48] V.D. Blondel, J.-L. Guillaume, R. Lambiotte and E. Lefebvre, Fast unfolding of communities in large networks, *Journal of statistical mechanics: theory and experiment* **2008**(10) (2008), 10008.

[49] C. Sarasua, S. Staab and M. Thimm, Methods for intrinsic evaluation of links in the web of data, in: *European Semantic Web Conference*, Springer, 2017, pp. 68–84.

[50] W. Shen, J. Wang and J. Han, Entity linking with a knowledge base: Issues, techniques, and solutions, *IEEE Transactions on Knowledge and Data Engineering* **27**(2) (2015), 443–460.

Notes

[1] http://swse.deri.org/entity/

[2] https://github.com/cgueret/LinkedData-QA

[3] http://bit.ly/Linked-QA

[4] https://github.com/cgueret/LinkedData-QA/tree/master/reports

[5] https://km.aifb.kit.edu/projects/btc-2011/

[6] http://nl.dbpedia.org:8080/TripleCheckMate/

[7] https://github.com/AKSW/TripleCheckMate/releases/tag/DBpediaCampaign

[8] http://nl.dbpedia.org:8080/TripleCheckMate/

[9] http://people.aifb.kit.edu/mac/DBpediaQualityAssessment/experiments.html

[10] http://people.aifb.kit.edu/mac/DBpediaQualityAssessment/experiments.html

[11] http://oaei.ontologymatching.org/2010/im/

[12] http://dbtune.org/bbc/peel/

[13] http://skipforward.opendfki.de/wiki/DBTropes

[14] https://dws.informatik.uni-mannheim.de/en/research/rapidminerlodextension/

[15] http://ls3.rnet.ryerson.ca/predicatefinder/category/

[16] https://www.dropbox.com/s/m24xoxzm0h60ywl/correct.tar.gz?dl=1

[17] https://github.com/firmao/CEDAL

[18] https://github.com/dice-group/CEDAL/blob/master/100Sample.tsv

[19] https://drive.google.com/file/d/0B3W6K8QxmfLnc3FwV1lZd1hVemM/view

[20] https://github.com/criscod/SeaStar

[21] https://github.com/criscod/SeaStar/tree/master/data/plots

[22] http://lod-a-lot.lod.labs.vu.nl/

[23] http://sameas.cc/

[24] https://github.com/raadjoe/LOD-Community-Detection

[25] https://sameas.cc

[26] https://github.com/raadjoe/LOD-Community-Detection