MetaLink: A Travel Guide to the LOD Cloud

Wouter Beek[1], Joe Raad[2], Erman Acar[2], and Frank van Harmelen[2]

[1] Triply Ltd. `https://triply.cc` `wouter@triply.cc`
[2] Knowledge Representation & Reasoning Group, Vrije Universiteit Amsterdam
`https://krr.cs.vu.nl` {`j.raad,erman.acar,frank.van.harmelen`}`@vu.nl`

Abstract. Graph-based traversal is an important navigation paradigm for the Semantic Web, where datasets are interlinked to provide context. While following links may result in the discovery of complementary data sources and enriched query results, it is widely recognized that traversing the LOD Cloud indiscriminately results in low quality answers. Over the years, approaches have been published that help to determine whether links are trustworthy or not, based on certain criteria. While such approaches are often useful for specific datasets and/or in specific applications, they are not yet widely used in practice or at the scale of the entire LOD Cloud. This paper introduces a new resource called *MetaLink*. MetaLink is a dataset that contains metadata for a very large set of `owl:sameAs` links that are crawled from the LOD Cloud. MetaLink encodes a previously published error metric for each of these links. MetaLink is published in combination with LOD-a-lot, a dataset that is based on a large crawl of a subset of the LOD Cloud. By combining MetaLink and LOD-a-lot, applications are able to make informed decisions about whether or not to follow specific links on the LOD Cloud. This paper describes our approach for creating the MetaLink dataset. It describes the vocabulary that it uses and provides an overview of multiple real-world use cases in which the MetaLink dataset can solve non-trivial research and application challenges that were not addressed before.

Keywords: Semantic Web · Linked Open Data · Identity Management · Graph Navigation

1 Introduction

The ability to follow links between datasets is perhaps the most important theoretic benefit of Linked Open Data. The following of links in order to learn more about a data item is laid down in the fourth Linked Open Data rule [2] and it is the fifth star of Linked Open Data [10]. Unfortunately, in practice it is widely recognized that traversing the LOD Cloud indiscriminately may result in following incorrect links. Since the validity of an entire dataset can be jeopardized by following such incorrect links, LOD clients are often hesitant to follow links at all. The fear of following bad links undermines the basic purpose of Linked Data: the reuse of other people's datasets and the interpretation of a data item in the context of other people's assertions about that same item.

Over the past decade various approaches have been published that help determine whether links are trustworthy or not based on certain criteria. While such approaches are often useful for specific datasets and/or in specific applications, they are not yet widely used by clients in practice. The reason for this is that existing identity resolution approaches are relatively complex to implement, computationally expensive to use, make assumptions that are valid for some but not all datasets, and rely on properties like text labels and/or ontological axioms that are present in some but not all datasets. As such, existing identity resolution approaches are inherently at odds with graph-based navigation clients, which are generally light-weight, run on commodity hardware (e.g., within a web browser), and are expected to be so generic as to be able to navigate the entire LOD Cloud, or at least a significant subset of it.

This paper introduces *MetaLink*, a new resource that helps light-weight clients navigate the links of LOD Cloud-sized graphs. MetaLink is a dataset that contains metadata for a very large set of `owl:sameAs` links that are crawled from the LOD Cloud. It encodes a previously published error metric for each of these identity links and also publishes the grouping of links in terms of the originally asserted equivalence sets as well as in terms of so-called communities that are the result of an existing clustering algorithm. As such, MetaLink provides detailed metadata about the trustworthiness of specific identity links, as well as an overview of high-trust links for specific nodes in the LOD Cloud.

MetaLink is a meta-dataset that contains *metadata* about `owl:sameAs` assertions that have been published publicly. As such, MetaLink only becomes truly useful when combined with *data* that contains nodes that are described in MetaLink. For example, we will use LOD-a-lot, a dataset that is based on a crawl of a very large subset of the LOD Cloud. By combining MetaLink and LOD-a-lot (or any other Linked Dataset that uses terms that appear in the LOD Cloud), applications are able to make informed decisions about whether or not to follow specific links on the LOD Cloud. This results in multiple real-world use cases in which the MetaLink dataset can be used to solve non-trivial research and application challenges that were not addressed before.

This paper makes the following contributions:

1. A specification of the requirements for a meta-dataset of identity links.
2. An approach for generating MetaLink, a meta-dataset of identity links that follows these requirements.
3. An implementation of the approach that is able to generate instances of MetaLink in a repeatable, low-cost, and scalable way.
4. Illustrations of use cases that are enabled by the availability of MetaLink.

The rest of this paper is structured as follows: Section 2 gives the motivation for creating MetaLink, discusses related work, and provides a list of design requirements. In Section 3, the approach for generating, storing and querying MetaLink is described. The implementation of MetaLink is described in Section 4. Some of the uses that are enabled by the availability of MetaLink are presented in Section 5. Section 6 concludes the paper.

2 Motivation

Graph-based traversal is an important navigation paradigm for the Semantic Web[3]. The basic idea behind Linked Data is that datasets are not only semantically described on an individual basis, but also they are interlinked with one another. Indeed, the use of links in order to interconnect datasets is specified by the fourth and last Linked Data rule [2]: *"Include links to other URIs, so that [data clients] can discover more things"*. As such, the creation of links is more than a courtesy sign of Linked Data etiquette. Links are necessary in order to express the full meaning of a dataset. Full meaning is achieved by positioning formally described nodes in the context of the wider fabric of meaning that is asserted by the ever expanding Web of Data. This essential semantic step of contextualizing a dataset by connecting it to the global fabric of meaning, is also known as the fifth star of Linked Open Data [2]: *"Link your data to other data to provide context"*.

The formal correlate of the practice of linking is specified in the Web Ontology Language (OWL) by the `owl:sameAs` predicate [12]. This predicate denotes the identity relation (i.e., the smallest equivalence relation). Had the Semantic Web been an isolated Knowledge Representation system, there would have been no need for an identity-denoting predicate in the first place. Indeed, in such a closed system each distinct concept could have been expressed by a distinct name, and that would have lifted the need for any kind of linking (such knowledge representation systems are said to adhere to the Unique Name Assumption (UNA)). But the Semantic Web is not an isolated system, it is a world-wide collaborative effort that already includes hundreds of thousands of datasets that are specifically intended to be interpreted and used in the context of each other.

While Linked Open Data *theory* focuses on the necessity to traverse links in order to interpret the meaning of data within a wider context, in *practice* it is widely recognized that traversing the LOD Cloud indiscriminately may result in following incorrect links and – by combining Linked Data that maybe should not have been combined – that may result in low-quality answers.

Let us take a concrete example. Suppose that we are traversing the LOD-a-lot dataset, starting out with the DBpedia IRI `dbr:President_Barack_Obama`. By following an `owl:sameAs` link we reach the Freebase IRI `fb:m.05b6w1g`, and from there we follow another `owl:sameAs` link to reach another DBpedia IRI: `dbr:Barack_Obama_cabinet`. We have only followed two identity links and we

[3] In this paper we use the following RDF prefix declarations for brevity:

- `dbc`: `http://dbpedia.org/resource/Category:`
- `dbr`: `http://dbpedia.org/resource/`
- `fb`: `http://rdf.freebase.com/ns/`
- `owl`: `http://www.w3.org/2002/07/owl#`
- `rdfs`: `http://www.w3.org/2000/01/rdf-schema#`
- `skos`: `http://www.w3.org/2004/02/skos/core#`
- `meta`: `https://krr.triply.cc/krr/sameas-meta/def/`

are already in big semantic trouble! We are now conflating a person who is an important member of a group with the entirety of that group[4].

While the notion of providing context by following links should be the main benefit of using Linked Data, the validity of an entire dataset can be jeopardized by following only one incorrect link. As a result of this extremely high cost of following one single potentially erroneous link, Linked Data clients are hesitant in following links at all. This is unfortunate, because a plethora of valid `owl:sameAs` links can be followed into a vast number of possibly relevant datasets, encapsulating potentially useful information.

2.1 Related work

Over time, an increasing number of studies in Semantic Web have shown that the identity predicate is used incorrectly for various reasons (e.g. heuristic entity resolution techniques, lack of suitable alternatives for `owl:sameAs`, context-independent classical semantics). This misuse has resulted in the presence of a number of incorrect `owl:sameAs` statements in the LOD Cloud, with some studies estimating this number to be around 2.8% [11] or 4% [17], whilst others suggesting that possibly one out of five `owl:sameAs` in the Web is erroneous [9].

Some vocabularies have proposed alternatives to `owl:sameAs` with different or no semantics. For example, `umbel:isLike` statements denote similarity instead of identity and are symmetric but not transitive; `skos:exactMatch` statements are symmetric and transitive, but indicate "a high degree of confidence that the concepts can be used interchangeably across a wide range of information retrieval applications," which is semantically very different from the notion of identity. As a result, the semantics of the closure that is calculated over this variety of statements is unclear.

Various approaches have been proposed for detecting erroneous identity statements, based on the similarity of textual descriptions associated to a pair of linked names [5], UNA violations [14, 20], logical inconsistencies [11, 16], network metrics [8], and crowd-sourcing [1]. However, existing approaches either do not scale in order to be applied to the LOD Cloud as a whole, or they make assumptions about the data that may be valid in some datasets but not in others (we refer the reader to [19] for more details). For example, in the LOD Cloud not all names have textual descriptions, many datasets do not include vocabulary mappings, or they lack ontological axioms and assertions that are strong enough to yield inconsistencies. While all of the here mentioned approaches for erroneous identity links detection are useful in some cases, this paper presents a solution that can be applied to all datasets of the entire LOD Cloud.

[4] Notice that such conflations are generally allowed in natural language semantics, where policies enacted by the Obama administration are commonly denoted by phrases like "Obama's policies".

2.2 Requirements

While a large number of identity resolutions approaches exist, such approaches are relatively complex to implement, computationally expensive to use, make assumptions that are valid for some but not all datasets, and rely on properties like text labels and/or ontological axioms that are present in some but not in all datasets. As such, existing identity resolution approaches are inherently at odds with graph-based navigation clients, which are generally light-weight, run on commodity hardware (e.g., within a web browser), and are expected to be so generic as to be able to navigate the entire LOD Cloud, or at least a significant subset of it. Since light-weight LOD clients can already be assumed to implement basic Linked Data querying mechanisms like SPARQL or Linked Data Fragments (LDF) [21], it makes sense to publish a solution to the identity resolution problems in the form of a Linked Open Dataset. Such an identity meta-dataset must meet the following requirements in order to be truly usable for a wide variety of LOD clients:

1. **Scalable.** The approach for generating the identity meta-dataset must be applicable on a very large scale. This requirement is needed in order to be able to apply the here presented approach on an increasingly larger scale, ultimately at the scale of the entire LOD Cloud.
2. **Reliable.** The metric that indicates the trustworthiness or error degree of identity links must be good enough to be relied upon in many client applications. This requirement is a trade-off with respect to Requirement 1: since the meta-dataset must be applicable on the scale of the LOD Cloud, it cannot extensively rely on dataset-specific features.
3. **Ordered.** It is often interesting to know the order in which an identity between two terms has been asserted. For example, even though formal semantics states that identity assertions are entirely symmetric, in practice most linkset publishers put their own terms in the subject position and the terms they link to in the object position.
4. **Modular.** An identity meta-dataset must be able to integrate with existing datasets. It must not put an unnecessary burden on the client that wishes to use it, but must tap into the dataset that the client is already using.
5. **Standards-compliant.** An identity meta-dataset must be encoded using open standards[5]. This allows light-weight clients that already implement LOD standards to interpret and process the identity meta-dataset with relatively small implementation changes.
6. **Broadly applicable.** It must be possible to use the identity meta-dataset in order to achieve a broad range of research goals and applications that cannot be achieved (or very difficult to achieve) by existing means.
7. **Low-cost.** Since it is very difficult to sustain resources within an academic setting, the cost of generating, hosting, and using the identity meta-dataset must be very low. Specifically, it must be much lower than the traditional approach of loading all the dataset into a (memory-intensive) triple store and/or processing all data in memory.

[5] `https://www.w3.org/standards/`

3 MetaLink data model

This section details the data model of MetaLink, an identity meta-dataset that implements the requirements specified in Section 2.2. Figure 1 gives an overview of the MetaLink vocabulary, and Figure 2 shows an example of two identity assertions together with their corresponding MetaLink metadata.

3.1 Implicit & explicit identity assertions

MetaLink distinguishes between two types of identity statements: those that are explicitly asserted (Definition 1), and those that can be derived from such explicit assertions through entailment (Definition 2).

Definition 1 (Explicit Identity Relation). *The explicit identity relation for an RDF graph G is represented by the tuple $\langle V, E, w\rangle$. E is the set of directed edges $\{(x, y) \mid \langle x, \texttt{owl:sameAs}, y\rangle \in G\}$. V is the set of vertices $\{x \mid (x, y) \in E \lor (y, x) \in E\}$. $w : E \to \{1, 2\}$ is the weight function:*

$$w((x,y)) := \begin{cases} 1 & \textit{if } (y,x) \notin E \\ 2 & \textit{if } (y,x) \in E \end{cases}$$

The order in which assertions have been made (Requirement 3) is preserved by reifying explicit identity assertions using the properties `rdf:subject` and `rdf:object`. While there is some overhead in also asserting the predicate term (`rdf:predicate`) for each identity assertion, doing so keeps the application of the RDF vocabulary recognizable (Requirement 5), while at the same time opening up the possibility for storing links that do not use the `owl:sameAs` property in the future.

Definition 2 (Implicit Identity Relation). *The implicit identity relation for an RDF graph G is represented by the tuple $\langle V', E'\rangle$. E' is the closure of E under equivalence (reflexivity, symmetry, transitivity). V' is the set of vertices $\{x \mid (x, y) \in E' \lor (y, x) \in E'\}$.*

While it is essential to store explicit identity assertions, this is not the case for implicit identity assertions. Firstly, assertions that only belong to the implicit identity relation follow from the explicit identity relation in systematic ways, i.e., according to OWL entailment rules. An identity meta-dataset can rely on the same systematicity in order to derive metadata about implicit assertions from the recorded metadata about explicit assertions. Secondly, the implicit identity relation is impractically large to store. In general, an identity set of size N can be expressed by $N-1$ explicit identity assertions, but the corresponding closure contains N^2 implicit identity assertions. Since identity sets can contain tens of thousands of terms, the difference between the implicit and the explicit identity relation for one identity set can already amount to billions of assertions.

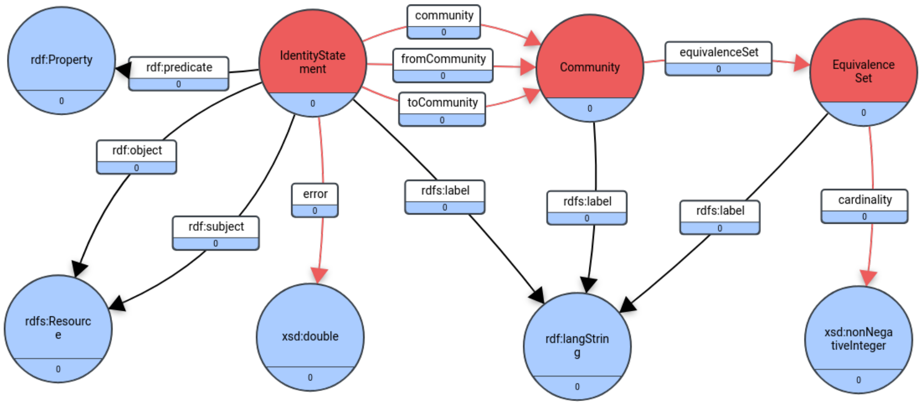

Fig. 1. Vocabulary of the MetaLink dataset. Classes are displayed by circles and properties are displayed by arcs. The MetaLink-specific classes and properties are displayed in red, the blue classes and properties are reused from existing vocabularies. The vocabulary can be accessed at https://krr.triply.cc/krr/metalink/graphs.

3.2 Singleton & non-singleton equivalence sets

The implicit identity relation assigns exactly one equivalence set to every term (Definition 3). The set of all equivalence sets forms a partition of the domain of discourse V'. Because MetaLink only records explicit identity links, it also only records non-singleton equivalence sets.

Definition 3 (Equivalence set). *For a specific term x, the corresponding equivalence set is* $[x]_\sim := \{y \mid (x,y) \in E'\}$.

3.3 Communities

In order to implement the scalability and reliability requirements (Requirements 1 and 2), MetaLink uses the community detection approach for identity links that is introduced in Raad et al. [18]. This approach uses the Louvain algorithm in order to cluster every connected component of the explicit identity relation into one or more communities. Communities partition equivalence sets, which partition the domain of discourse. Once the communities have been detected, an error metric is calculated (Section 3.4). This results in the only identity metric that has been calculated at the required scale and that has acceptable accuracy. In addition, this metric is calculated by an efficient, low-cost algorithm (Requirement 7). MetaLink distinguishes between explicit identity assertions that form intra-community links and ones that form inter-community links (Definition 4). MetaLink uses the `:community` property to relate identity assertions to the communities to which their subject and object terms belong. The subproperties `:fromCommunity` and `:toCommunity` are used to relate inter-community links to their respective communities. MetaLink uses the `:equivalenceSet` property

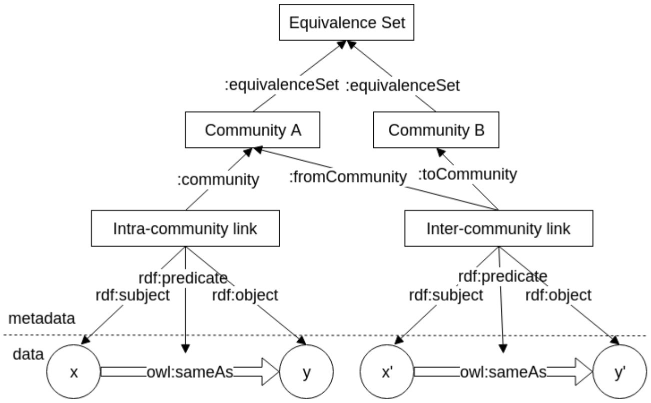

Fig. 2. An example of two `owl:sameAs` assertions (below the dotted line) in combination with the corresponding MetaLink annotations (above the dotted line).

in order to relate communities to their corresponding equivalence sets (see the top half of Figure 2).

Definition 4 (Intra- and Inter-Community Links). *An intra-community link is an edge* $(x, y) \in E$ *such that* x *and* y *belong to the same community. An inter-community link is an edge* $(x, y) \in E$ *such that* x *and* y *belong to different communities.*

3.4 Error metric

After detecting the community structure in each equivalence set, an error degree for each identity link is computed. This error degree, described in details in [18], is computed based on two elements: (a) the density of the community for intra-community links or the density of both communities for inter-communtiy links, and (b) whether the link is reciprocally asserted ((x, y) and (y, x)). Overall, reciprocally asserted links have a lower error degree than non-reciprocally asserted identity links. Furthermore, links that belong to more densely connected communities are more likely to be correct. This results in an error degree for each identity link ranging from 0.0 (most likely correct) till 1.0 (most likely incorrect). The experiments in [18] show that indeed the higher an error degree of an identity link is, the more likely it is erroneous. Specifically, the manual evaluation conducted by the authors in [18] show that links with error degree >0.99 are in most cases erroneous (∼1M identity links), whilst identity links with error degree <0.4 are in most cases correct (∼400M identity links). MetaLink uses the `:error` property to store the error degree.

3.5 Separation between metadata and data

In line with the modularity requirement (Requirement 4), MetaLink makes a clean separation between data and metadata. The data (displayed below the dotted line in Figure 2) is intended to be delivered by the data consumer, either up-front or during the process of online link traversal.

The relationship between data and metadata is established with the RDF reification properties (`rdf:subject`, `rdf:predicate`, and `rdf:object`). The reification properties clearly communicate to data consumers that they are traversing the boundary between data and metadata. Notice that it would have been possible to establish links between terms in the data (`x`, `x'`, `y`, and `y'` in Figure 2), but doing so would have made the distinction between metadata and data less noticeable to a modest data consumer.

The link assertion on the left-hand side ($\langle x, \texttt{owl:sameAs}, y\rangle$) is an example of an intra-community link, so the generic `:community` property is used to relate (the subject and object terms in) the identity link to Community A. The link assertion on the right-hand side ($\langle x', \texttt{owl:sameAs}, y'\rangle$) is an example of an inter-community link, so the more specific `:fromCommunity` and `:toCommunity` properties are used to relate (the subject and object terms in) the identity link to Communities A and B. Both communities have the same equivalence set (property `:equivalenceSet`).

4 Implementation

MetaLink is created based on the TSV file[6] published as a part of [18]. This TSV file contains rows for over 330M non-reflexive `owl:sameAs` assertions that are drawn from the LOD-a-lot dataset [6]. The TSV file has the following columns:

- The subject or object term, whichever comes lexicographically first.
- The subject or object term, whichever comes lexicographically last.
- The calculated error degree: a value between 0.0 (probably correct) and 1.0 (probably incorrect).
- The weight of the link: 2 if the symmetric link also appears in LOD-a-lot, and 1 if this is not the case.
- A unique identifier for the equivalence set to which the link belongs.
- The cardinality of the equivalence set.
- Either a unique identifier for a community (for inter-community links), or a pair of from/to (in that order) unique community identifiers (for intra-community links).

The TSV file is used as the input for the MetaLink creation script. Because the order of the terms within links is relevant in MetaLink (Requirement 3), we use the original LOD-a-lot file in order to determine the order for each row in the TSV file. The script is written in SWI-Prolog that has extensive support for RDF, and is publicly available[7]. The script generates an N-Triples file that contains 4,352,602,452 unique triples and describes 556,152,454 non-reflexive `owl:sameAs` links.

[6] `https://krr.triply.cc/krr/sameas/assets/5c16733d68c97e02a691c19a`

[7] `https://github.com/wouterbeek/sameas_script`

Class	# instances
`meta:IdentityStatement`	556,152,454
`meta:Community`	55,697,160
`meta:EquivalenceSet`	48,999,148

Property	# triples
`meta:cardinality`	48,999,148
`meta:community`	410,706,139
`meta:equivalenceSet`	55,697,160
`meta:error`	556,152,454
`meta:fromCommunity`	145,446,315
`meta:toCommunity`	145,446,315

Table 1. Overview of the size of the composition of the MetaLink dataset in terms of its classes and properties.

4.1 HDT: low-cost usage

In order to implement the low-cost requirement (Requirement 7) we cannot publish the MetaLink dataset in a traditional triple store. Even though there are triple stores that are able to store 4.3B triples, such services are relatively costly to set up. Also, MetaLink is only truly useful when combined with a dataset in which the identity metadata can be used. Since we want people to use the MetaLink meta-dataset in the context of the LOD-a-lot dataset, it would be preferable to expose MetaLink together with the 28.3B LOD-a-lot triples. For this reason we create a Header Dictionary Triples (HDT) [7] file. HDT provides a popular low-cost alternative to memory-intensive Linked Data publication approaches. By working almost exclusively from disk, HDT allows the MetaLink meta-dataset and the LOD-a-lot dataset to be queried from commodity hardware such as a regular consumer laptop. Table 1 shows statistics about the MetaLink classes and properties that are obtained from the HDT file. The MetaLink HDT file and its index (36GB each) are published at persistent URI with a citable DOI:

- MetaLink HDT file (`https://doi.org/10.5281/zenodo.3227976`)

Since it has been made available online in April 2019 as part of the Zenodo Linked Data and Semantic Web communities, this dataset has attracted[8] more than 161 views (131 unique), 42 downloads (15 unique), and a number of tweets by members of the Semantic Web community.

4.2 TriplyDB: low-cost hosting

MetaLink and LOD-a-lot are published in a TriplyDB[9] instance over at `https://krr.triply.cc`:

- MetaLink (`https://krr.triply.cc/krr/metalink`)
- LOD-a-lot (`https://krr.triply.cc/krr/lod-a-lot`)
- MetaLink with LOD-a-lot (`https://krr.triply.cc/krr/lod-a-lot-plus`)

[8] Statistics collected by Zenodo and visible on the dataset's web page

[9] `https://triply.cc`

TriplyDB is an HDT-based Linked Data hosting platform. Human users can navigate the MetaLink and LOD-a-lot datasets with an HTML-based browser. Machine users can use a Linked Data Fragments (LDF) API (Requirement 5).

5 Use cases

This section briefly describes five concrete use cases for which MetaLink is an enabler. While we do not have the space here to expand on these use cases in great detail, they do show the impact and utility of MetaLink for academic research and LOD client applications (Requirement 6).

5.1 Follow-Your-Nose

In Section 2, we saw that performing a Follow-Your-Nose approach quickly resulted in following incorrect links such as the following:

```
fb:m.05b6w1g owl:sameAs dbr:President_Barack_Obama .
```

A light-weight Linked Data client typically does not have a module that can estimate the trustworthiness of links. However, such a client is probably able to query the MetaLink dataset with the following query:

```
select ?err {
  [ rdf:subject fb:m.05b6w1g;
    rdf:object dbr:President_Barack_Obama;
    :error ?err ]. }
```

For example, this query can be performed with the Comunica SPARQL engine (`http://comunica.linkeddatafragments.org/`) by using the MetaLink Triple Pattern Fragments API as a backend (`https://api.krr.triply.cc/datasets/krr/metalink/fragments`). The result for `?err` is 1.0, in other words: most likely an incorrect link. Based on this information, a client may choose to not follow this link.

5.2 Question Answering

The Follow-Your-Nose use case can be extended to cover queries of arbitrary complexity. We will illustrate this based on two SPARQL queries from the literature. The first question is “Who are the band members of ABBA?”, which appears in Buistra et al. [3] as the following SPARQL query:

```
select distinct ?member ?label {
  ?member
    skos:subject dbc:ABBA_members;
    rdfs:label ?label.
  filter(lang(?label) = ‘en’)}
```

In order to follow identity links into the LOD Cloud, we change this into the following query:

Result	≤ 1.0	≤ 0.8	≤ 0.6	≤ 0.4	≤ 0.2	≤ 0.0
Björn Ulvaeus (band member)	28	8	8	3	2	2
Agnetha Fältskog (band member)	26	4	4	2	1	1
Anni-Frid Lyngstad (band member)	9	3	3	2	1	1
Benny Andersson (band member)	6	2	2	1	1	1
Ola Brukert (drummer)	3	2	2	1	1	1
Agnetha Ulvaeus (Agnetha F. married name)	2	0	0	0	0	0
Stig Andersson (band manager)	9	4	4	1	1	1
Gert van der Graaf (stalker of Agnetha Fältskog)	2	0	0	0	0	0
Benny Anderssons Orkester (new band)	5	3	3	0	0	0
Stig Andersson (sportsman)	2	2	2	0	0	0

Table 2. Results of the ABBA band member query using different error degrees in MetaLink.

```
select distinct ?member ?label {
  ?member
    owl:sameAs*/skos:subject/owl:sameAs* dbc:ABBA_members;
    rdfs:label/owl:sameAs* ?label.
  filter(lang(?label) = 'en')}
```

Table 2 shows the number of results for different error degrees in MetaLink. The column under ≤ 1.0 shows the results when all available links are followed, i.e., without distinguishing between high and low error degrees. These results include the four correct answers (display in the first four rows), offering many alternative names/IRIs from DBpedia, Wikidata, OpenCyc, New York Times, and other datasets. The table also shows that there are many results that may be considered incorrect, like the drummer of ABBA, the manager, and stalker of one of the ABBA band members. The subsequent columns lower the error degree, resulting in more trustworthy links. The use of MetaLink for this query is inconclusive: the number of incorrect results decreases, but the number of alternative names for the correct results decreases too.

Our second question is “Through which countries does the Yenisei river flow?” which appears in Lopez et al. [13] as the following SPARQL query:

```
select distinct ?uri ?string {
  dbr:Yenisei_River dbp:country ?uri.
  optional {
    ?uri rdfs:label ?string.
    filter(lang(?string) = 'en')}}
```

When this query is performed with error degree ≤ 0.3, the two correct answers are returned: Russia and Mongolia. When the error degree is above 0.3, more than 30K results are returned, including hundreds of unrelated geographic places, the concept of creative writing, and the mythical creature Gorgon. For this query it is clear that the LOD Cloud contains incorrect links that destroy the value of following links, and that MetaLink can be used to circumvent this risk.

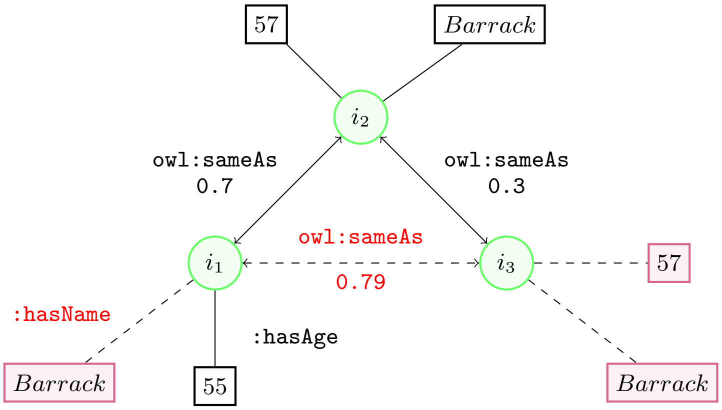

Fig. 3. An example of fuzzy reasoning over the error degrees in MetaLink. Solid edges denote explicit assertions; dashed lines denote implicit assertions. The derived error degree is calculated with t-conorm f_p. The predicted properties are displayed in red boxes.

5.3 Fuzzy reasoning

In Section 2.1, we saw that there have been ample attempts at replacing the role of `owl:sameAs` with less strict alternatives that denote various shades of relatedness. Unfortunately, such alternative linking properties fail to oil the wheels of Semantic Web when they seek to replace potentially faulty identity links with links that have no semantics whatsoever. Since MetaLink assigns a specific error degree between 0.0 and 1.0 to each `owl:sameAs` link, it can be used in order to assign a fuzzy alternative to the classical binary OWL semantics.

For example, by borrowing the notion of *t-conorm* from Fuzzy Logic [15] we can assign fuzzy error degrees to the implicit (i.e., missing) identity statements. A t-conorm is a function $f : [0,1] \times [0,1] \rightarrow [0,1]$ that is commutative, monotonic, associative, and that treats 0.0 as the identity element. As such, t-conorm is often used as the fuzzy correspondence of the binary logic operator $\vee$.

We now give two examples of t-conorms in the context of the MetaLink metadataset. Firstly, the standard maximum t-conorm is defined as $f_{\max}(a,b) := \max(\{a,b\})$. Intuitively, $f_{\max}$ adopts a pessimistic perspective on errors; considering all explicit identity links with known error degree, it assigns the maximum error degree to the corresponding implicit identity link. This perspective is useful if we are entirely skeptical about the truth values of the implicit or missing identity links. The downside of that perspective is that, it results in a less diverse set of truth values, since larger values are copied throughout the graph. Secondly, the probabilistic sum t-conorm, defined as $f_p(a,b) := a + b - a \cdot b$, results in the assignment of more diverse fuzzy truth values to implicit identity links.

Figure 3 shows an example of three instances that belong to the same equivalence set. Solid lines denote explicit `owl:sameAs` statements. MetaLink error degrees 0.7 and 0.3 are associated with the links (i_1, i_2) and (i_2, i_3), respectively. The implicit link (i_1, i_3) is shown with a dashed line and red text: its error degree 0.79 is calculated using t-conorm f_p. If the maximum t-conorm were used instead, this error degree would have been 0.7. Due to monotonicity, both t-conorms assign a derived error degree that is at least as high as the maximum of the two explicit error degrees. This reflects an important intuition: the derived identity link (i_1, i_3) cannot be more trustworthy than either of the explicit identity links that it is based on.

Another application scenario is the prediction of properties for missing and/or conflicting properties. MetaLink allows such predication to be applied to entities that belong to the same equivalence set. Figure 3 shows an example of predicted property values that is based on the existing properties in combination with the error degrees in MetaLink. Initially i_1 is missing `:hasName`, which is completed based on the information in i_2. Moreover, i_3 is initially missing both `:hasName` and `:hasAge`. For the latter, there is a conflict with respect to the age value (i_1 has value 55 but i_2 has value 57), and priority is given to the value that is associated with the equivalent entity that has the lower error degree.

5.4 Erroneous Identity Link Detection

The error degrees attributed to each `owl:sameAs` link in MetaLink is computed based on the recent approach by Raad et al. [18]. In this work, the authors showed that when the threshold is fixed at 0.99 (i.e. links higher than this threshold are considered erroneous), the approach enables detection of a large number of erroneous `owl:sameAs` (93% recall). However, the evaluation also shows that a number of correct identity links were attributed to such a high error degree (precision between 40% and 73%). As a consequence, correct links with such high error degree would be also discarded from applications aiming at using a higher quality subset of the LOD cloud, hence leading to the unwanted loss of additional information. Therefore, one possible and direct use case would be to apply more computationally expensive approaches to this smaller subset of `owl:sameAs` links for minimizing this information loss. Most importantly, since MetaLink is published in combination with LOD-a-lot, these approaches can rely on additional information besides the `owl:sameAs` network topology, in which these error degrees were computed from.

5.5 Erroneous Identity Link Benchmarking

In recent years, a number of approaches aiming at detecting erroneous identity links were introduced. Such approaches tend to make certain trade-offs, either by leveraging scalability over the accuracy of the approach [11, 14, 20], or the other way around [4, 16, 5]. These two categories of approaches are traditionally applied to different datasets, with the former generally applied to large real-world datasets (e.g., DBpedia), whilst the latter usually applied to smaller, mostly synthetic datasets (e.g., subset of links from the Ontology Alignment Evaluation Initiative OAEI). In addition, results generated from such approaches (e.g., the erroneous/correct links detected/validated by the approach, their error/confidence score, the dataset, the manually evaluated gold standard by the authors) become hardly accessible after publication due to several technical and social factors. As a consequence, the current situation shows that these results are hardly reproducible and comparable in practice. MetaLink can be deployed as a platform for solving this problem. Firstly, it allows both categories of approaches to be tested on the same dataset, where less scalable approaches can be tested on a subset of these links. Secondly, the vocabulary of MetaLink can be extended in a way that allows different approaches to publish their error degree and manually evaluated links. This will allow approaches to be directly compared and deployed long past the publication of their results.

6 Conclusion & Future work

This paper has presented MetaLink, an identity meta-dataset that stores the error degree of a large number of `owl:sameAs` statements that occur in the LOD Cloud. The availability of such an error degree is valuable, especially for light-weight Linked Data clients that currently do not have alternative means for determining the validity of identity links. The MetaLink approach is complementary to existing identity resolution approaches that may be more accurate, but that are not (yet) published for the scale of the LOD Cloud. We have presented several use cases for which MetaLink is an enabler, including question/answering systems, error link detection/benchmarking, and research into alternative identity semantics. The version of MetaLink presented in this paper is based on the data collected from the LOD Laundromat 2015 crawl. Since the construction of this dataset is completely automated, an updated version will be published as soon as a new crawl of the LOD Cloud is made available.

References

1. Acosta, M., Zaveri, A., Simperl, E., Kontokostas, D., Auer, S., Lehmann, J.: Crowdsourcing linked data quality assessment. In: International Semantic Web Conference. pp. 260–276. Springer (2013)
2. Berners-Lee, T.: Linked data-design issues (2011), `http://www.w3.org/DesignIssues/LinkedData.html`
3. Buikstra, A., Neth, H., Schooler, L., Teije, A.t., Harmelen, F.v.: Ranking query results from linked open data using a simple cognitive heuristic. In: IJCAI-11 (2011)
4. CudreMauroux, P., Haghani, P., Jost, M., Aberer, K., De Meer, H.: idmesh: graph-based disambiguation of linked data. In: International conference WWW. pp. 591–600. ACM (2009)
5. Cuzzola, J., Bagheri, E., Jovanovic, J.: Filtering inaccurate entity co-references on the linked open data. In: International DEXA Conference. pp. 128–143. Springer (2015)
6. Fernández, J.D., Beek, W., Martínez-Prieto, M.A., Arias, M.: Lod-a-lot. In: International Semantic Web Conference. pp. 75–83. Springer (2017)

7. Fernández, J.D., Martínez-Prieto, M.A., Gutiérrez, C., Polleres, A., Arias, M.: Binary RDF representation for publication and exchange (HDT). Web Semantics: Science, Services and Agents on the World Wide Web **19** (2013)
8. Guéret, C., Groth, P., Stadler, C., Lehmann, J.: Assessing linked data mappings using network measures. In: Extended Semantic Web Conference. pp. 87–102. Springer (2012)
9. Halpin, H., Hayes, P.J., McCusker, J.P., McGuinness, D.L., Thompson, H.S.: When owl:sameAs isn't the same: An analysis of identity in Linked Data. In: International Semantic Web Conference. pp. 305–320. Springer (2010)
10. Hausenblas, M.: 5 ⋆ open data (2012), `http://5stardata.info/`
11. Hogan, A., Zimmermann, A., Umbrich, J., Polleres, A., Decker, S.: 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**, 76–110 (2012)
12. Horrocks, I., Patel-Schneider, P.F., Harmelen, F.V.: From SHIQ and RDF to OWL: The making of a web ontology language. Web semantics: science, services and agents on the World Wide Web **1**(1), 7–26 (2003)
13. Lopez, V., Unger, C., Cimiano, P., Motta, E.: Evaluating question answering over linked data. Web Semantics: Science, Services and Agents on the World Wide Web **21**, 3–13 (2013)
14. de Melo, G.: Not quite the same: Identity constraints for the web of linked data. In: AAAI. AAAI Press (2013)
15. Novák, V., Perfilieva, I., Mockor, J.: Mathematical principles of fuzzy logic, vol. 517. Springer Science & Business Media (2012)
16. Papaleo, L., Pernelle, N., Saïs, F., Dumont, C.: Logical detection of invalid sameas statements in rdf data. In: International Conference EKAW. pp. 373–384. Springer (2014)
17. Raad, J.: Identity Management in Knowledge Graphs. Ph.D. thesis, University of Paris-Saclay (2018)
18. Raad, J., Beek, W., Van Harmelen, F., Pernelle, N., Saïs, F.: Detecting erroneous identity links on the web using network metrics. In: International semantic web conference. pp. 391–407. Springer (2018)
19. Raad, J., Pernelle, N., Saïs, F., Beek, W., van Harmelen, F.: The sameas problem: A survey on identity management in the web of data. arXiv preprint arXiv:1907.10528 (2019)
20. Valdestilhas, A., Soru, T., Ngomo, A.C.N.: Cedal: time-efficient detection of erroneous links in large-scale link repositories. In: International Conference on Web Intelligence. pp. 106–113. ACM (2017)
21. Verborgh, R., Vander Sande, M., Hartig, O., Van Herwegen, J., De Vocht, L., De Meester, B., Haesendonck, G., Colpaert, P.: Triple pattern fragments: a low-cost knowledge graph interface for the web. Web Semantics: Science, Services and Agents on the World Wide Web **37**, 184–206 (2016)