DBpediaSameAs: an Approach to Tackle Heterogeneity in DBpedia Identifiers

Andre Valdestilhas
AKSW, Department of Computer Science
Augustusplatz 10
D-04109 Leipzig, Germany
valdestilhas@informatik.uni-leipzig.de

Natanael Arndt
AKSW, Department of Computer Science
Augustusplatz 10
D-04109 Leipzig, Germany
arndt@informatik.uni-leipzig.de

Dimitris Kontokostas
AKSW, Department of Computer Science
Augustusplatz 10
D-04109 Leipzig, Germany
kontokostas@informatik.uni-leipzig.de

ABSTRACT

The DBpedia dataset has multiple URIs within the dataset and from other datasets connected with (transitive) `owl:sameAs` relations and thus referring to the same concepts. With this heterogeneity of identifiers it is complicated for users and agents to find the unique identifier which should be preferably used. We are introducing the concept of DBpedia Unique Identifier (DUI) and a dataset of linksets relating URIs to DUIs. In order to improve the quality of our dataset we developed a mechanism that allows the user to rate and suggest links. As proof of concept an implementation with a graphical web user interface is provided for accessing the linkset and rating the links. The DBpedia sameAs service is available at `http://dbpsa.aksw.org/SameAsService`.

Categories and Subject Descriptors

M.0 [**Knowledge Management**]: Knowledge Acquisition; H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Semantic Web, Linked Data, DBpedia, SameAs, Link Rate

1. INTRODUCTION

As DBpedia [3] was evolving during the 9 years of its existence, the community extended the linksets to DBpedia resources. Thus, DBpedia has more than one URI that represents the same resource, which leads to the identifier heterogeneity problem. For instance a DBpedia resource can contain `owl:sameAs` links to other data sets such as FreeBase[1], Wikidata, GeoNames[2] or yago[3].

Further DBpedia has more than one URI representing the same resource within the dataset, e.g. the `dbpedia:Brassil`[4] has at least the following equivalents within the DBpedia `dbpedia:Republica_Federativa_do_Brasil`, `dbpedia:ISO_3166-1:BR` and `dbpedia:Brazil` which are all redirecting to `dbpedia:Brazil`. Thus a problem to consider is to directly resolve any of the equivalents directly to the final URI e.g. `http://dbpedia.org/resource/Brazil` without any redundancies.

Also, according to Halpin et. al. [1] and Wood et. al. [4], *sameas.org* has collected millions of triples with `owl:sameAs` relations. It would be important to promote reciprocal owl:sameAs confirmation mechanisms and develop effective trust mechanisms to assure the quality of `owl:sameAs` relations.

To tackle the identifier heterogeneity problem we are making the following contributions:

- We describe an approach for the mitigation of the identifier heterogeneity problem and implement a prototype where the user is able to evaluate existing links, as well as suggest new links to be rated.
- The ability to generate statistics about good and bad links which, brings the possibility to have a quality control for the links to DBpedia.
- We define the DBpedia Unique Identifier (DUI), which instead of several transient `owl:sameAs` DBpedia URIs for the same final address, now is possible to have a unique URI from DBpedia. A DUI goes directly to the final address instead of having to process several possible intermediate results. For example, with a URI from *Freebase*, 17 redundant URIs from DBpedia where avoided or if one used a service such as sameAs.org, 1141 URIs would be avoided.

The rest of the paper is organized as follows: section 2 represents a proposed approach for tackling the identifier heterogeneity problem, we evaluate our work in section 3, in section 4 we focus on related work, and finally section 5 concludes the paper and outline future work.

[1] Freebase project webpage: `http://freebase.com`

[2] GeoNames project and exploration webpage: `http://geonames.org`

[3] Yago project webpage:`http://yago-knowledge.org`

[4] Throughout the paper we are using the following namespace definitions: `owl`: `http://www.w3.org/2002/07/owl#`, `dbpedia`: `http://dbpedia.org/resource/`.

2. REPRESENTATION OF THE IDEA

This section provides an explanation about our main idea, such as implementation and descriptions.

Before continuing the work, there are some definitions that were adopted.

- **Normalization of the URI**: Is understood by normalizing URIs, the fact of eliminating redundancies.
- **DBpedia unique identifier**: The DBpedia Unique Identifier (DUI) is an unique URI that identifies a resource in the DBpedia repository and also is the result of our normalization.

The idea started with a stand alone service on the web that solves the problem where the user provides a URI as parameter and instead of several transient URIs with `owl:sameAs` property, the user receives a single DUI from our service.

2.1 The work-flow

The work-flow for requesting the DUI of a given resource is represented in fig. 1. Firstly, the user will provide a URI from some address, i.e. FreeBase. Then, instead of possible several results of URIs with the property `owl:sameAs`, our system will return a DUI. Consequently, the user has a possibility to rate, verify, validate, and suggest a different link. Then the rate can give us a chance to have statistics about the quality of the links.

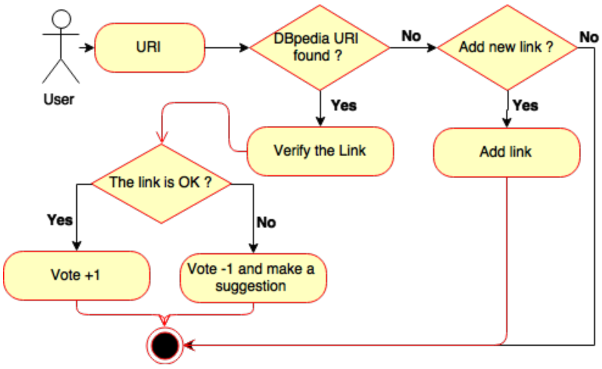


Figure 1: General work-flow.

A service, also was implemented, where the user can provide a URI and the API will return the DBpedia identifier like a URI that represents the `owl:sameAs` about the URI provided.

2.2 Methodology

This section describes in four steps the technique and how the idea was developed, from phase of importing links to a relational database until the development of the service on the web and a GUI.

(1) The files with triples that contains `owl:sameAs` links, were downloaded. (2) All triples were imported in a relational database [5], because we will use some characteristics of a relational database i.e. comparative with voting system in future works. (3) An implementation of a service on the web was provided, where the user enters the URI and receives a DUI. (4) In order to provide an interface to access this service were created a web system that receive as input a URI, return as output an DBpedia identifier and allow rate and make suggestions about the resulting link.

Figure 2 presents the relation of the contribution in a graph form.

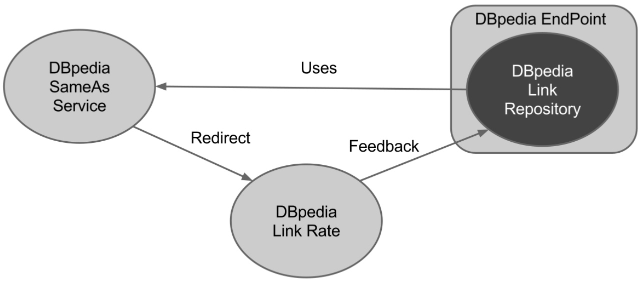


Figure 2: Relation of the contributions.

Where the DBpedia Link Repository uses the DBpediaSameAs service in order to tackle the heterogeneity and giving the appropriate DUI, that redirects the user to the DBpedia Link Rate interface, thus, providing a feedback to the DBpedia Link Repository, therefore, improving the quality of the DBpedia endpoint.

3. EVALUATION

The aim of this qualitative[6] evaluation was centered in verifying the behavior of the service DBpediaSameAs, the Graphical User Interface (GUI) that gives the possibility to verify and rate the links.

There are chosen 3 evaluation criteria:

(1) **Normalization on DBpedia URIs**: With this criteria was evaluated if the DBpediaSameAs can provide an normalization on DBpedia URIs. (2) **Rate the Links**: Where was evaluated if the DBpediaSameAs can provide a way to rate the links. (3) **DBpediaSameAs as service**: Was evaluated if DBpediaSameAs can provide a stand alone service on the web that brings the normalization on DBpedia URIs.

3.1 Normalization on DBpedia URIs

The criteria used in this evaluation are uniquely to tackle heterogeneity, that was observed during the search of co-references between different data sets with a problem about redundancies.

When was used a URI from freebase in order to obtain a DBpedia URI was observed that at least 3 URIs were returned, that drives to the same final address.

As an example of a real case, executed in our public server, with a URI from Freebase:

[5] `http://tinyurl.com/creatdb`

[6] `http://tinyurl.com/rmethod`

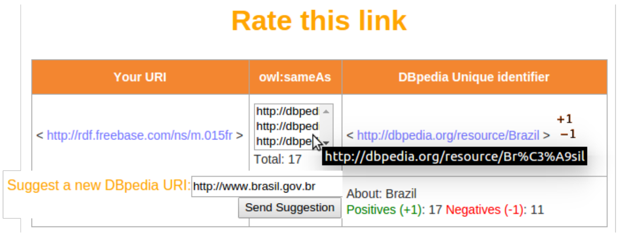
```
$ curl http://dbpsa.aksw.org/¬
  SameAsService/SameAsServlet?uris=http%3¬
  A%2F%2Frdf.freebase.com%2Fns%2Fm.015fr

returns: http://dbpedia.org/resource/¬
  Brazil
```

Where, in this case, instead of 17 URIs from DBpedia, that goes to the same final address, our approach drives the user directly to the final address.

As can be observed on the figure 4 that approach the transitive and redirect URIs, where show that with this approach instead of have several URIs the user can have only one from the DBpediaSameAs. Thus, in this way, providing a normalization on DBpedia URIs.

3.2 Rate the links

In order to have a link rating, were implemented a GUI that allows the users to give some feedback, suggestions, in this way, improving the quality of the links. The rate is a quite simple process, the GUI just ask the user to rate the link with +1 if the link attends your expectations or -1 if the link is wrong or some type of spam. The GUI was developed using concepts from prefix.cc [7] and work from Zaveri[5] such our system of rate (+1 and -1) and the standard of the web documents. Some improvements and personalization, also was provided, such as the suggestions and the possibility to check the link. The figure 3 shows the moment when the user clicked on the -1 and indicated that the user didn't like the link and was asked to make a suggestion of a new URI.



Rate this link

Your URI	owl:sameAs	DBpedia Unique identifier
< http://rdf.freebase.com/ns/m.015fr >	http://dbped http://dbped http://dbpe Total: 17	< http://dbpedia.org/resource/Brazil > +1 −1

http://dbpedia.org/resource/Br%C3%A9sil

Suggest a new DBpedia URI: http://www.brasil.gov.br

About: Brazil
Positives (+1): 17 Negatives (-1): 11

Figure 3: Rate a link. Available at `http://dbpsa.aksw.org/SameAsService`

The field about a suggestion for a new link will only appear when the user are not satisfied with the current link, then, when clicking on the -1, then the system will ask for an optional suggestion.

3.3 Results

The results of this work could also be expressed in numbers that was obtained during importing triples to the relational database and with some results from the sameAs.org web site. A total of 62,531,487 triples imported into our database, the time was 2,220 seconds for the whole operation, thus, was noticed that 28,167 triples were imported per second. The source code used to obtain the results is available in our github repository[8].

3.3.1 Transitive and Redirect Links

Transitive and Redirect Links are redundancies at DBpedia that supposed has a link to the same place, in other words, they use `owl:sameAs` property, this links will redirect another links, will provide a transition between the links, that's why the name transitive. In this case, instead of using this transitive links that points to the same final destination URI, this final destination URI will be used directly. The figure 4 try to make more clear this explanation.

Was discovered and treated 6,473,988 triples with transitive and redirect links from 62,531,487 imported links among 142 domains inside DBpedia. Then, 10.35% of the links can be avoided in some cases.

3.4 Discussion

The DBpediaSameAs was evaluated with its normalization of URIs, link rate, and DBpediaSameAs as a stand alone service on the web. As results of the normalization a DUI was obtained in order to tackle the heterogeneity. In other words, instead of several URIs e.g. from sameAs.org one DUI was obtained. The link rate functionality further allows to improve the quality of the dataset.

Despite, the GUI of DBpediaSameAs, also a stand alone service on the web was developed that brings the functionality to get a DUI without a GUI for agents or people which don't need to use the DBpediaSameAs in a Graphical mode, allowing use as an off-the-shelf component.

4. RELATED WORK

The work [5] elaborates a data quality assessment methodology in DBpedia, which comprises of a manual and semi-automatic process. This work drive us to a reinforcement about the concept of data quality used in our work, when in our case will be more a manual process and also we are able to improve the DBpedia data quality.

The work [2], presents a two staged experiment for the creation of gold standards that act as benchmarks for several interlinking algorithms. The similar aspects of this works are: The validation of links and a dubbed manual validation, where the user i.e. validator or evaluator specifies whether a link generated by an interlinked tool is correct or incorrect. The results of the link validation process are used to learn presumably better link specifications and thus achieving high-quality. Also, this work proposes an experiment to investigate the effect of user intervention in dataset interlinking on small knowledge bases.

4.1 A related problem with sameAs.org

The sameAs.org is a service that leading source of co-reference data on the Semantic Web. For example, when the web site sameAs.org is accessed with a URI from Freebase that should bring information about a country called Brazil.

Was used the URI (`http://rdf.freebase.com/ns/m.015fr`) as parameter to the service, and is received as return more than 1140 URIs as shown in fig. 5, but the user can have a doubt about which one is the correct.

Our work is not an alternative to the sameAs web site, but brings possibilities, like, was noticed that the sameAs.org

[7] `http://prefix.cc`

[8] `https://github.com/firmao/dbpedia-links/blob/master/CreateDB.sh`

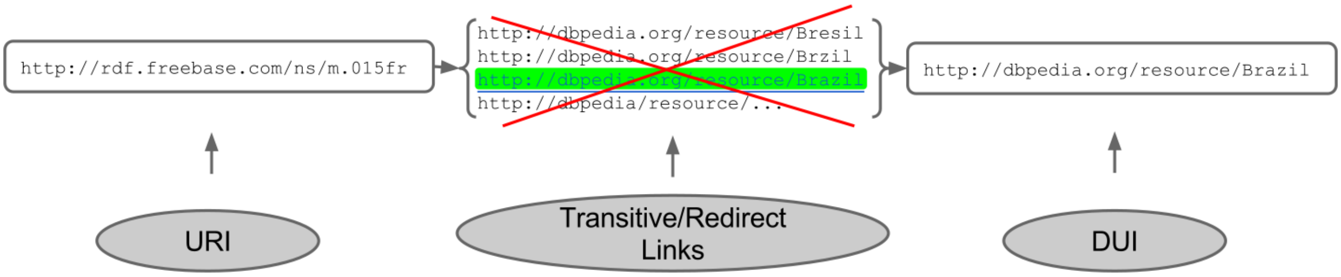
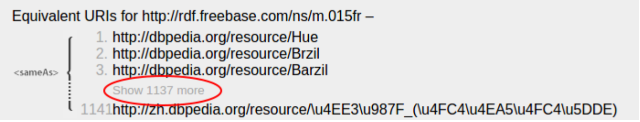

http://rdf.freebase.com/ns/m.015fr
http://dbpedia.org/resource/Bresil
http://dbpedia.org/resource/Brzil
http://dbpedia.org/resource/Brazil
http://dbpedia/resource/...
http://dbpedia.org/resource/Brazil
URI
Transitive/Redirect Links
DUI

Figure 4: Transitive / Redirect links in DBpedia.



Equivalent URIs for http://rdf.freebase.com/ns/m.015fr –

<sameAs>

1. http://dbpedia.org/resource/Hue
2. http://dbpedia.org/resource/Brzil
3. http://dbpedia.org/resource/Barzil

Show 1137 more

1141 http://zh.dbpedia.org/resource/\u4EE3\u987F_(\u4FC4\u4EA5\u4FC4\u5DDE)

Figure 5: Several URIs with the property `owl:sameAs` from the web site sameAs.org.

does not provide a way to rate the link, but with this rating, is possible to improve the quality of the data, and bring some facility to the user.

5. CONCLUSION AND FUTURE WORKS

An approach was provided to tackle the heterogeneity working with `owl:sameAs` redundancies that were observed during researching co-references between different data sets and providing a unique DBpedia identifier and give the chance to rate the resulting links and make suggestions.

A proof of concept was implemented as a computer web system in order to present and validate our idea and every concept of this work. The source code is available [9].

Was noticed in our results that there are benefits when a considerable number of `owl:sameAs` redundancies can be avoided. Rating the links allow users to make link suggestions brings more quality to the repository, and the stand alone service on the web allow you to use the DBpediaSameAs also in a command line textual environment and can be used as an off-the-shelf component.

For the future we plan to: (1) make a study about the results of link rating. This needs a period of usage of the DBpediaSameAs service in order to gather sufficient results for proper analysis. (2) An implementation case with more members of the DBpedia community. A study about how will be the behavior when implement with the DBpedia community.

6. ACKNOWLEDGMENT

We would like to acknowledge, National Council for Scientific and Technological Development (CNPq) [10] and Universität Leipzig for their support. Special thanks to Markus Ackermann for a essential help with the deployment of application and very good suggestions. Additionally,research activities of this paper were funded by grants from the EU's 7th & H2020 Programmes for projects ALIGNED (GA 644055), GeoKnow (GA 318159) and LIDER (GA 610782).

7. REFERENCES

[1] 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 P. F. Patel-Schneider, Y. Pan, P. Hitzler, P. Mika, L. Zhang, J. Z. Pan, I. Horrocks, and B. Glimm, editors, *International Semantic Web Conference (1)*, volume 6496 of *Lecture Notes in Computer Science*, pages 305–320. Springer, 2010.

[2] M. Hassan, J. Lehmann, and A.-C. N. Ngomo. Interlinking: Performance assessment of user evaluation vs. supervised learning approaches. In *24th International World Wide Web Conference (WWW 2015): workshop: Linked Data on the Web (LDOW2015), Florence, Italy, May 18 to 22, 2015, Proceedings*, 2015.

[3] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer. DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, 2014.

[4] D. Wood, M. Zaidman, L. Ruth, and M. Hausenblas, editors. *Linked Data: Structured data on the Web.* Manning, 2014.

[5] A. Zaveri, D. Kontokostas, M. A. Sherif, L. Bühmann, M. Morsey, S. Auer, and J. Lehmann. User-driven quality evaluation of dbpedia. In *Proceedings of the 9th International Conference on Semantic Systems*, I-SEMANTICS '13, pages 97–104, New York, NY, USA, 2013. ACM.

[9] `https://github.com/firmao/DBPediaLinkSameAs.git`

[10] http://www.cnpq.br/