Class and Instance Equivalences in the Web of Linked Data: Distribution and Graph Structure

Salvador Sanchez-Alonso¹(✉), Miguel A. Sicilia¹, Enayat Rajabi², Marçal Mora-Cantallops¹, and Elena Garcia-Barriocanal¹

¹ University of Alcalá, Alcalá de Henares, Spain
{salvador.sanchez,msicilia,marcal.mora,elena.garciab}@uah.es
² Cape Breton University, Sydney, Canada
enayat.rajabi@cbu.ca

Abstract. The Web of Linked Open Data (LOD) is a decentralized effort in publishing datasets using a set of conventions to make them accessible, notably thought RDF and SPARQL. Links across nodes in published datasets are thus critical in getting value for the LOD cloud as a collective effort. Connectivity among the datasets can occur through these links. Equivalence relationship is one of the fundamental links that connects different schemas or datasets, and is used to assert either class or instance equivalence. In this article, we report an empirical study on the equivalences found in over 59 million triples from datasets accessible via SPARQL endpoints in open source data portals. Metrics from graph analysis have been used to examine the relationships between repositories and determine their relative importance as well as their ability to facilitate knowledge discovery.

Keywords: Linked Open Data · Equivalence · RDF · SPARQL

1 Introduction

The cloud of Linked Open Data (LOD) is the result of the effort of different institutions or individuals in publishing schemas and data openly based upon a set of conventions including RDF as the fundamental information sharing model and following the philosophy of lack of centralized storage or control of the World Wide Web. Similar to the Web, the value of LOD depends critically on links across nodes which support aggregating or contrasting information at different levels. Among all the possible semantics that links may bear, equivalence is of special importance, as it provides a basis for merging either schemas or concrete data, depending on the resources being asserted as equivalent.

Class equivalence axioms entail replaceability of two classes (or in general, of class expressions), which in practical terms when considering graphs of linked data results in the possibility of merging graphs that may reside at different
nodes. Further, this provides an extension for all the asserted instances of the classes. Thus, those equivalences are extremely important in interoperability and schema reuse of data in the LOD cloud, so they deserve special attention. Instance equivalence assertions have a scope limited to pairs of individuals, and complement class assertions with the important benefit of theoretically allowing for aggregating data on particular entities. We have left other declarative expressions as property equivalences or declaring that two individuals are different for further work.

In spite of the practical importance of equivalence relationships, existing studies suggest that they are not in widespread use in the LOD cloud, and are not consistently used to frame the semantics of the data exposed by nodes in an attempt to maximize opportunities for merging or fusing with other nodes. In this paper, we report preliminary results for a systematic study of equivalences across the Web of Data, as a first step towards a more complete account of the topic. Results show that general purpose repositories such as DBpedia or Geonames play a central role in the LOD cloud, acting as a bridge for other repositories to be linked to the other datasets, which in turn provides better possibilities for knowledge discovery to data consumers.

The rest of this paper is structured as follows. Section 2 provides an overview of the axioms that enable interlinking in the LOD repositories and briefly surveys previous studies. Section 3 details the procedure carried out to collect the data that is later analyzed in Sect. 4. Finally, conclusions are provided in Sect. 5.

2 Background

Links across datasets are a key element in the LOD cloud. Arguably, one of the most valuable types of relationships between entities in the LOD datasets is equivalence. We take here the notions of equivalence that appear in the OWL2 language.

An equivalent classes axiom $\text{EquivalentClasses}(CE_1 \ldots CE_n)$ states that all of the class expressions $CE_i$, $1 \leq i \leq n$, are semantically equivalent to each other. This axiom allows one to use each class as a synonym for one other. In other words, in any expression in the ontology containing such an axiom, any $CE_i$ can be replaced with $CE_j$ without affecting the meaning of the ontology. This has strong implications. In the case of asserting the equivalence of two primitive classes in different datasets, this entails that the graphs can be merged and taken together.

Similarly, an individual equality axiom $\text{SameIndividual}(a_1 \ldots a_n)$ states that all of the individuals $a_i$, $1 \leq i \leq n$, are equal to each other. In consequence, their names should in theory be used interchangeably.

The $\text{owl:sameAs}$ axiom has been subject to different studies starting from those of Halpin et al. [1]. The problems identified with the use of $\text{sameAs}$, and the different uses have been known for more than a decade now, but they appear to persist [2,3]. [4] also identified several issues involving the $\text{owl:sameAs}$ property in a linked data context including merging assertions from sources with different
contexts, and the need to explore an operational semantic distinct from the strict logical meaning provided by OWL. The sameas.cc dataset [5] is a large collection of identity assertions extracted from the Web of Data. However, we are more interested in the graph structure of equivalences in this article, as it happens when traversing linked data.

3 Data Collection

The point of departure for data collection was checking the SPARQL endpoints available in CKAN that are used to generate LOD cloud diagrams.

Concretely, we systematically checked each of the nodes appearing in datahub\(^1\) to discover the existing live SPARQL endpoints of datasets. Out of 711 Linked Open Data datasets with SPARQL endpoints, 93 endpoints were responding to queries. The examination of availability of each endpoint was performed by writing a simple Python script using the SPARQLWrapper\(^2\).

It should be noted that this procedures may have two important limitations. On one hand, not necessarily every published dataset in the LOD cloud is registered and curated at that repository. On the other hand, the availability of SPARQL endpoints may vary over time [6], therefore, the obtained sample cannot be guaranteed to be comprehensive, but at least it represents a snapshot in time of collecting datasets from the LOD cloud.

Equivalences were obtained by using a simple SPARQL query as follows:

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
SELECT ?s ?p ?o
WHERE { ?s ?p ?o.
    FILTER (?p IN ( owl:equivalentClass , owl:sameAs ) ) }
```

It was observed that some endpoints returning high round figures of results, which might be attributed to some constraints on the query engine for large result sets. In order to check that possible behaviour, SPARQL queries aggregating the number of results with \(\text{COUNT}(\ast)\) were used.

4 Results and Discussion

4.1 About Node IDs

According to the W3C definition\(^3\), a blank node ID, also known as “blank node” or “bNode”, is a local identifier used in some concrete RDF syntaxes or RDF store implementations. A node ID is not, therefore, any special type of data record but instead a way to refer to a certain node within the confines of a given

---

1. https://old.datahub.io.
2. https://pypi.org/project/SPARQLWrapper/.
3. https://www.w3.org/TR/rdf11-concepts/.
file or knowledge-base: outside of the file where the node ID appears, the node is unnamed and thus cannot be linked or referred.

In our study, a total of 329 triples containing node IDs (0.3% of the total of 123,387 equivalent class axioms) were found in 11 repositories. The Brazilian Politicians dataset, for instance, includes triples like the following:

http://purl.org/ontology/mo/Arranger,equivalentClass,nodeID://b19867

This triple states that an equivalence between the class Arranger and an unnamed resource b19867 exists. The use of node IDs is handy to identify things via e.g. an inverse functional property in a knowledge-base when we do not want to give a reasoner on the file excess work. At the same time, unnamed nodes are heavily used in rdf:List structures. Given that the Brazilian politicians repository, in our example, includes seven blank nodes declared to be equivalent to the class http://purl.org/ontology/mo/Arranger, we assume that a list of seven resources, whose names are unknown to the outside knowledge-bases, are “Arrangers”. In the same way, seven Listeners and seven Conductors are listed among others. We also understand that all such resources have a value for an inverse functional property, which can be used to refer to the nodes from outside the knowledge-base.

The complete distribution of triples including node IDs per repository is shown in Table 1:

| Repository | # nodeIDs |
|------------|-----------|
| AEMET metereological dataset | 9 |
| Allie Abbreviation and Long Form Database in Life Science | 18 |
| Brazilian politicians | 44 |
| DBpedia in Japanese | 11 |
| DBpedia in Portuguese | 44 |
| Library of the Chilean national congress | 3 |
| GeoLinkedData | 6 |
| Open Data from the Italian National Research Council | 6 |
| Dbnary (Wiktionary as Linguistic Linked Open Data) | 148 |
| UniProt RDF schema ontology | 7 |
| Environment data UK | 1 |

4.2 Equivalent Classes

A total of 123,387 equivalent class axioms were found in the data collected from the responsive datasets (see Sect. 3). Out of a total of 59,383,718, this number represents only the 0.2% of the collected triples.
Inspecting the subjects of the predicates, the most frequent one was FOAF 0.1 Person with 96 occurrences. Other highly referenced FOAF classes are Agent, Image and Document. Other frequent objects above 20 occurrences include a number of classes in the DBPedia ontology among others such as Person, Organization, Event, Work, Place, Language, City or Country. Two highly ranked classes in the http://www.openlinksw.com/schemas/ namespace were Tweet and User.

Looking at the objects, the most frequent concept is Schema.org, including CreativeWork and Person with 61 and 60 occurrences respectively. Other Schema.org classes above 25 occurrences are ImageObject and Product. Occurrences above 25 among others are Person in DBPedia and FOAF, Agent in Dublin Core, UserAccount and Post in SIOC, and Person in PIM datasets.

Looking at the endpoints, a number of DBPedia instances had between 400 and 1,500 equivalences (pt, ja, de, fr) and the global http://dbpedia.org/sparql amounted for 10,000 (maybe capped by the SPARQL query). The URI Burner4 data service was the most frequent amounting to 100,000 records. Given that this endpoint is an aggregation service, it can be considered as a special case.

Although other perspectives were possible, the analysis of relationships between different repositories was prioritized as the most interesting possible study. Therefore, we proceeded to extract the relationships between repositories to further analyse the data. The procedure implied the creation of a file with relationships between the repositories: for all triples including the equivalent class axiom, source and target repositories were recorded. The extraction was carried out in the following steps:

- In step 1, a Python script extracted the URLs of the two repositories, removed the name of the resources, and obtained the rest of the URL.
- In step 2, human experts cleaned the data by removing duplicates and fixing errors to later identify the repository to which each URL belonged.

With this file as an input, we built a weighted directed graph where the weight assigned to a given relationship between repository A and repository B depended on the number of classes in A related by equivalent class axioms to the other classes in B, duplicates excluded. A first version of the graph was created including 96 nodes and 214 relationships. This graph was later revised to transform node IDs (see previous section) to self links, remove the isolated nodes and small graphs. We finally obtained the “giant component” (which is a finite connected fraction of the entire graph’s vertices), shown in Fig. 1, and consisted of only 41 nodes and 125 relationships. Knowledge-bases like AEMET meteorological dataset, Datos.bcn.cl, Environment Agency Bathing Water Quality, Geo Vocab, Semantic Sensor Network Ontology, Lotico or MindsWap, which had no links to the other repositories, were filtered out.

In Fig. 1, the node size has been adapted according to the in-degree value – i.e. the number of incoming relations from other repositories– of each repository. Not surprisingly, general purpose knowledge-bases such as DBPedia or common

4 http://linkeddata.uriburner.com/.
use ontologies and schemes such as W3C, Dublin Core, the BIBO ontology or Schema.org are among the most referenced nodes. In the same Figure, 4 communities were identified and coloured in green, purple, orange and blue. According to the algorithm used [7], two repositories belong to the same cluster (community) if they are more densely connected together than to the rest of the network, so this was the criterion to cluster nodes together in a community.

4.3 Same-as Relationships

In the data collected from the responsive datasets, a total of 59,260,331 sameAs axioms were found out of a total of 59,383,718 (this represents a 99.8% of all triples collected). The endpoints with larger number of axioms are the British National Library (∼2M), the Isidore scientific collection [5] (∼1M), the OpenLink Virtuoso public endpoint and the Portuguese DBPedia (∼200K each) and the Allie Database of abbreviations in science and the French DBPedia (∼100K each).

The distribution of the subjects and objects is relatively flat, with only elements from the Last.fm endpoint having more than 300 occurrences, in which appears to be a specificity of the internals in that node.

Following a similar procedure to the one described for equivalent classes, a weighted directed graph including 275 nodes and 491 relations was built. The giant component (again, removing disconnected nodes and sub-graphs) was 269

---

5 [https://isidore.science/](https://isidore.science/).
nodes and 485 relationships. The most referenced repositories were DBPedia, Geonames, the Library of Congress, and other repositories that are either general purpose or a reputed source of information for a given field. Although one could expect the fact that DBPedia has a higher in-degree value due to incoming links from external knowledge-bases, it is not so clear how DBPedia in minority languages such as Sanskrit, Yoruba, Yiddish, Zhongwen, Wallon or Volapük—to name just a few—can attract large number of links from the other repositories.

Figure 2 shows how the repositories in the full graph can be classified into five communities. The biggest one—including most DBpedias in many different languages—contains 72% of the nodes, while the rest contain 15% (ontologies and semantic Web institutions), 5% (libraries), 4.73% and 2.55%. In the Figure, node size is proportional to the in-degree value of each repository.

If only those nodes with relations with four or more repositories are chosen—the 4-core graph—we have what is shown in Fig. 3. This graph, where the node size is proportional to the repository in-degree value—i.e. the number of references from external repositories—, shows the more referenced repositories according to the number of relationships. As one would expect, general purpose knowledge-bases such as DBpedia, Geonames, and Freebase are among the most central ones.

According to the Linked Open Data model, it is important for a repository to be linked to as many other repositories as possible, as this promotes new
knowledge discovery. Betweenness centrality, which is a Social Network Analysis property that measures the extent to which a node lies on the paths between other nodes, is the most representative metric from our perspective to evaluate discoverability. Figure 4 shows the subset of repositories with a higher betweenness value, which in some way represents the datasets that provide better discovery pathways to their knowledge-base users. Perhaps not surprisingly, the repositories with higher betweenness centrality are also those with a higher number of outgoing links –out-degree metric in Social Network Analysis–. These are therefore the repositories whose resources distribute better their outgoing links, i.e. reference resources in a higher number of other repositories.
5 Conclusions and Outlook

This paper has described insights about class and instance equivalences from a sample of SPARQL endpoints in the cloud of Linked Open Data. Class equivalences show sparse connection patterns, with most frequent equivalences unsurprisingly from and to “upper concepts” as person, together with some technology specific concepts as user or tweet. Instance equivalences appear evenly distributed in terms of subjects and objects, suggesting a sparse distribution, but showing a more imbalance distribution considering endpoints.

General purpose repositories such as DBpedia, Geonames or Freebase are more central than the rest of repositories, as expected. This fact highlights that linking a dataset to these repositories opens better possibilities for knowledge discovery and it is highly beneficial for both dataset publishers and dataset consumers.

Properties like out-degree and betweennes centrality seem to be quite connected. This can lead us to the conclusion that the more varied datasets with high centrality we (as data creators) link our datasets, the more likely we are to be a connected to the other repositories on the Web. Also, betweenness centrality can be used as a metric in knowledge discovery to determine the flow of information in a knowledge graph.

References

1. 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: Patel-Schneider, P.F., et al. (eds.) ISWC 2010. LNCS, vol. 6496, pp. 305–320. Springer, Berlin, Heidelberg (2010). https://doi.org/10.1007/978-3-642-17746-0_20
2. Halpin, H., Hayes, P.J., Thompson, H.S.: When owl:sameas isn’t the same redux: towards a theory of identity, context, and inference on the semantic web. In: Christiansen, H., Stojanovic, I., Papadopoulos, G. (eds.) CONTEXT 2015. LNCS, vol. 9405, pp. 47–60. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-25591-0_4
3. 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)
4. Ding, L., Shinavier, J., Finin, T., McGuinness, D. L.: owl:sameAs and linked data: an empirical study. In: Proceedings of the 2nd Web Science Conference, April 2010 (2010)
5. Beek, W., Raad, J., Wielemaker, J., Van Harmelen, F.: sameAs.cc: the closure of 500M owl:sameAs statements. ESWC 2018. LNCS, vol. 10843, pp. 65–80. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93417-4_5
6. Rajabi, E., Sanchez-Alonso, S., Sicilia, M.A.: Analyzing broken links on the web of data: an experiment with DBpedia. J. Assoc. Inf. Sci. Technol. 65(8), 1721–1727 (2014)
7. Blondel, V., Guillaume, J., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech: Theory Exp. 2008, 10008 (2008)