Detection of Invalid Identity Links Statements in RDF Knowledge Graphs

Nathalie Pernelle[1], Joe Raad[1], and Fatiha Saïs[1]

Université de Paris-Sud, Laboratoire de Recherche en Informatique,
Bâtiment 650, F-91405 Orsay Cedex, France
`firstname.lastname@lri.fr`,
home page: `http://www.lri.fr`

Abstract. We are experiencing an unprecedented production of data, published online as *Linked Data*. In this context, `owl:sameAs` links are declared to express identity relation between resources that refer to the same real world entity. However, recent research discussions have shown issues in the use of `owl:sameAs`. In this workshop, we will present an overview of the symbolic and numerical approaches aiming to detect invalid `owl:sameAs` statements or to represent alternative identity links.

Keywords: `owl:sameAs`, Identity Link Invalidation, Alternative Links

1 Identity Problem

Identity is an old and thorny topic. According to Leibniz, resources that are identical are considered to share the same properties: with Ψ denoting the set of all properties, the 'Indiscernibility of Identicals' ($a = b \rightarrow (\forall_{\psi \in \Psi})(\psi(a) = \psi(b))$) states that resources that are identical share the same properties, and its converse, the 'Identity of Indiscernibles' ($(\forall_{\psi \in \Psi})(\psi(a) = \psi(b)) \rightarrow a = b$), states that resources that share the same properties are identical.

In the semantic web, identity statements allow to access to complementary descriptions of the same resource. However, several studies have shown that a `owl:sameAs` link is often used incorrectly in practice, and that they often express an identity that is context-dependent [2]. Since `owl:sameAs` is transitive, these incorrect statements can have wide-ranging effects in a large data graph like the LOD (Linked Open Data). The problem of the use of identity on the Semantic Web is now widely recognized, and referred to as the "sameAs problem" or the "Identity Crisis". This identity problem, has led to several discussions, and proposals for limiting its effects. In the workshop, we will present (1) an overview of the numerical and symbolic approaches that aim to detect erroneous identity links, and (2) some of the proposed alternatives for `owl:sameAs`.

2 Detection of invalid `owl:sameAs` and Alternative links

To discover invalid `owl:sameAs`, some approaches apply some reasoning step to detect that a `owl:sameAs` link leads to an inconsistent knowledge graph. For instance, [7] hypothesize that the datasets preserve the Unique Name Assumption

(UNA), and that violations of the UNA that can be detected once the transitive closure is applied may indicate that some of the involved links are erroneous. [4] detects inconsistencies by exploiting the semantics of 10 OWL 2 constructs such as AsymmetricProperty, or complementOf, while [5] is based on class disjointness, property mappings, (inverse) functional and local complete properties.

Other approaches compute a similarity score, or an error rate for each evaluated identity link. This score is either based on the similarity of the resources descriptions (e.g. property values) or on network metrics that can be computed for the identity graph (e.g. community structures [6], node centrality ...).

Some approaches have proposed alternative identity predicates that can be used to represent some of these erroneous links. [3] define a hierarchy of weaker predicates by their property of reflexivity, symetricity, and transitivity while other approaches aim to compute the semantic context (subpart of the ontology) in which a `owl:sameAs` can be considered as valid [1].

3 Conclusion

The approaches that aim to detect invalid `owl:sameAs` are based on ontology axioms, datasets characteristics (e.g. UNA, presence of differentFrom), similarities between property values, or network metrics. However, new approaches are needed to combine the advantages of all these approaches. Some alternative links have been proposed that can be used to represent erroneous but relevant context-dependent identity links. However, when a context is defined as a subpart of the ontology, such approaches need to be guided by expert knowledge. Indeed, the number of contexts can be very large when the ontology is complex.

References

1. Wouter Beek, Stefan Schlobach, and Frank van Harmelen. A contextualised semantics for owl: sameas. In *ISWC*, pages 405–419. Springer, 2016.
2. Li Ding, Joshua Shinavier, Tim Finin, Deborah L McGuinness, et al. owl: sameas and linked data: An empirical study. In *Web Science Conference*, 2010.
3. Harry Halpin, Patrick J Hayes, James P McCusker, Deborah L McGuinness, and Henry S Thompson. When owl:sameAs isn't the same: An analysis of identity in Linked Data. In *ISWC*, pages 305–320. Springer, 2010.
4. Aidan Hogan, Antoine Zimmermann, Jürgen Umbrich, Axel Polleres, and Stefan 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*, 10:76–110, 2012.
5. Laura Papaleo, Nathalie Pernelle, Fatiha Saïs, and Cyril Dumont. Logical detection of invalid sameas statements in rdf data. In *EKAW*, pages 373–384. Springer, 2014.
6. Joe Raad, Wouter Beek, Frank Van Harmelen, Nathalie Pernelle, and Fatiha Sais. Detecting erroneous identity links on the web using network metrics. In *ISWC*, 2008.
7. André Valdestilhas, Tommaso Soru, and Axel-Cyrille Ngonga Ngomo. Cedal: time-efficient detection of erroneous links in large-scale link repositories. In *Web Intelligence*, pages 106–113. ACM, 2017.