Identity Management in Knowledge Graphs

Joe Raad

University of Paris-Saclay, Paris, France

1 Background & Motivation

In a large and distributed knowledge space, such as the Web of Data, 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 in the Web of Data, it is unavoidable for this same real-world entity to be denoted by different names (IRIs, literals, blank nodes). Essential to the coherence of these large and geographically distributed knowledge graphs, publishers are encouraged to link their data. Such interlinking is mostly established by asserting that two names denote the same real-world entity. For this purpose, the W3C standardised knowledge representation language OWL introduced in 2004 the `owl:sameAs` predicate[1]. For instance, the statement $\langle dbr : Barack\_Obama$, `owl:sameAs`, $wdt : Q76\rangle$ indicates that both names from the *DBpedia* and *Wikidata* knowledge graphs refer to the same entity. With its strict logical semantics, this statement indicates that every property asserted to one name will also be inferred to the other. Hence, 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. With studies dating back to the early Semantic Web days showing that `owl:sameAs` is indeed misused in the Web [2, 7, 9], one can trace back their presence to several factors. From a philosophical point of view, the notion of 'Indiscernibility of Identicals' is problematic, since identity is actually context-dependent. For instance, 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). From a practical point of view, most `owl:sameAs` links are generated by entity resolution techniques, that employs practical strategies which are not guaranteed to be accurate. For instance, the precision of such tools ranged between 67% and 86% in the 2017 and 2018 Ontology Alignment Evaluation Initiative (OAEI)[2]. Finally, studies have shown that modellers have different opinions about whether two names refer to the same real-world entity or not. For instance in [7], 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

[1] `https://www.w3.org/TR/owl-ref#sameAs`

[2] `http://oaei.ontologymatching.org/2018/results/conference/index.html`

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.

Since suitable alternatives to `owl:sameAs` have yet to exist, or are rarely used in practice, a given Web application is forced to make a choice with respect to each `owl:sameAs` assertion it encounters. This problem of incorrect use of identity is not specific to the Web of Data, and is present in all Knowledge Representation systems [5, 11]. However, the problem is specifically alerting in the Web of Data due to its unprecedented size, the heterogeneity of its users and contents, and the lack of a central naming authority. By now, the problem of the identity use in the Semantic Web is widely recognised, and has been referred to as the "Identity Crisis" [1], and the "sameAs problem" [7]. As such, a proper approach towards the handling of identity links is required in order to make the Web of Data succeed as an integrated knowledge space.

2 Contributions

Identity management in knowledge graphs is the main objective of this thesis. It addresses one particular issue of the identity problem: the misuse of identity links in knowledge graphs. It does not cover related but distinct research topics such as entity resolution and ontology alignment, that focus on techniques [3] and frameworks [10] for establishing `owl:sameAs` links. In addition, this thesis does not address the historically significant 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 [6, 8]. This thesis investigates the use of `owl:sameAs` links in the Web of Data, and provides different, yet complementary solutions for this identity problem:

- Identity Management Service. In order to uncover different aspects over how identity is used in the Web of Data, and at the same time limit the technical burden for users in knowing whether two names refer to the same entity or not, this thesis introduces the `sameAs.cc` identity service. This Web service provides public access (query and download) to the largest collection of `owl:sameAs` statements that has been gathered to date from the Web of Data (over 558 million distinct `owl:sameAs`), and their resulting 49 million equivalence classes. Additionally, this thesis proposes an approach for efficiently computing and storing the equivalence classes, and presents different analyses over this data set of explicitly asserted `owl:sameAs` links, their closure, and their aggregation into namespaces. In contrary to this work's main predecessor [4], which considers a number of semantically different identity and similarity statements for computing the transitive closure, this resource provides semantically interpretable equivalence classes that can be used for instance by a DL reasoner in order to infer new facts.

- Approach for detecting erroneous identity links. With several studies showing that identity is incorrectly used in the Web of Data, there is an ever increasing need to detect these links to enhance the quality of knowledge graphs. For this, this thesis introduces a scalable approach for automatically detecting theses erroneous identity links, by making use of the `owl:sameAs` network's topology. Relying solely on the community structure of this network, and the symmetrical characteristic of the identity links, an error degree is calculated for each `owl:sameAs` statement. These error degrees are subsequently used for ranking these links, allowing potentially erroneous ones to be flagged, and potentially correct ones to be validated. Since the here presented approach is specifically developed in order to be applied to real-world data, the experiments are run on the `sameAs.cc` dataset. Based on a manual evaluation of around 1K `owl:sameAs` links, the results suggest that network measures are indeed effective in detecting erroneous `owl:sameAs` links (86% accuracy). The results also show that by solely removing 0.17% of the existing `owl:sameAs` links (the 1M `owl:sameAs` with an error degree higher than 0.99), the quality of the equivalence classes in the Web of Data can be significantly increased. Additionally, this thesis studies related works, and shows based on certain requirements (accuracy, scalability, type of assumptions presumed on the data), that the here introduced approach is currently the most effective technique for detecting erroneous `owl:sameAs` links in the context of the Web of the Data.

- A contextual identity relation. In many instances the classical interpretation of identity is too strong for particular purposes, and is not always required, as the notion of identity might change depending on the context. However, modelling all specific contexts in which an identity relation holds is cumbersome and, due to arbitrary reuse, and the Open World Assumption, it is not possible to anticipate all contexts in which an entity will be used. For this, this thesis introduces a new contextual identity relation. In this alternative notion, the contexts in which an identity relation between two class instances holds are explicit to the user with regard to a domain ontology. For defining the contextual identity relation, this thesis defines the notion of global contexts, their order relations, and the conditions that should be fulfilled for declaring an identity between two given instances in a certain (global) context. This thesis also introduces an algorithm for automatically detecting the most specific contexts in which a pair of instances are identical. Since not all contexts may be relevant (e.g. a context considering a measure's value without its unit), this algorithm can be guided by different sets of semantic constraints provided by experts for enhancing the detected contexts.

- Contextually linked knowledge graphs for life sciences. Cases in which objects can not be declared the same are quite common in scientific contexts, where experiments are mostly conducted by several scientists, in various circumstances, using similar but not the same products. This incapacity of semantically linking slightly different experiments has been a barrier for knowledge-based sys-

tems to fully exploit scientific data, as they are either weakly connected with little semantics (e.g. using `skos:closeMatch`), or are incorrectly declared the same (using `owl:sameAs`). In addition, the classical problems of the heterogeneity of the formats in which scientific data are published, and the terminological variations encountered across the multiple scientific datasets also remain important barriers in fully exploiting the large amount of data produced everyday. As a way for limiting these syntactic, semantic and identity problems, we adopt Semantic Web standards for introducing a new knowledge graph for life sciences. This graph is constructed in a mutual effort with domain experts from the French National Institute of Agricultural Research (INRA), describing two different domains: the mechanisms leading to the release of flavour compounds during food consumption, and the process of stabilisation of micro-organisms. As a way for semantically linking the different conducted experiments and their participants, we apply our approach for detecting the contexts in which such experiments can be considered the same. Additionally, we exploit the millions of detected contextual identity links in this scientific graph for discovering certain rules. These rules, when validated by the experts, can be used to predict with a certain degree of confidence, unobserved measures in the experiments, and consequently deployed for completing the constructed knowledge graph.

3 Conclusion & Perspectives

This thesis have investigated one specific research question: *how to limit the misuse of identity links in knowledge graphs*. For addressing this identity problem, this thesis analysed a number of existing related works that contributed to this research question (Chapter 2); proposed a resource that enables large-scale identity research (Chapter 3); showed the actual impact of misusing identity in the Web, whilst at the same time showing that it is possible to limit these adverse effects using solely the `owl:sameAs`'s network topology (Chapter 4); showed that the classical identity relation standardised in OWL is problematic, whilst introducing a new notion of context-dependent identity relation (Chapter 5) and showing its benefits in a newly constructed scientific knowledge graph (Chapter 6). Limitations of these contributions are discussed in Chapter 7, and a number of future directions are outlined.

The resources made available and the lessons learned from this thesis have motivated further extension of this work. For instance, an extension of the `sameAs.cc` resource have already emerged, in which each `owl:sameAs` link is reified for associating metadata, and several new higher-quality equivalence classes were made available to the community. These enhanced equivalences classes are generated based on specified error degrees, allowing Linked Data practitioners to control the trade-off between (a) using more links and benefiting from more contextual information from the Web, and (b) the risk of introducing erroneous information. In addition, this corpus of (enhanced) `owl:sameAs` links is currently being deployed for both question-answering and instance-based schema-alignments tasks. Finally, ongoing works are extending the here presented ap-

proach for computing the error degrees of `owl:sameAs` links, with the goal of improving its precision in detecting the erroneous ones. First work combines both network measures and outlier detection techniques for flagging potentially erroneous links, whilst the other work is investigating the use of mathematical programming and description similarity techniques for improving the here presented work, and provide higher quality equivalence classes.

References

1. Paolo Bouquet, Heiko Stoermer, and Daniel Giacomuzzi. Okkam: Enabling a web of entities. *I3*, 5:7, 2007.
2. Li Ding, Joshua Shinavier, Tim Finin, Deborah L McGuinness, et al. owl: sameas and linked data: An empirical study. In *Proceedings of the Second Web Science Conference*, 2010.
3. Alfio Ferrara, Andriy Nikolov, and François Scharffe. Data linking for the semantic web. *Semantic Web: Ontology and Knowledge Base Enabled Tools, Services, and Applications*, 169:326, 2013.
4. Hugh Glaser, Afraz Jaffri, and Ian Millard. Managing co-reference on the semantic web. In *Proceedings of the WWW Workshop on Linked Data on the Web, LDOW*, 2009.
5. John Grant and V. S. Subrahmanian. Reasoning in inconsistent knowledge bases. *IEEE Trans. Knowl. Data Eng.*, 7(1):177–189, 1995.
6. Harry Halpin. Sense and reference on the web (doctoral dissertation). University of Edinburgh, 2010.
7. 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 *International Semantic Web Conference*, pages 305–320. Springer, 2010.
8. Harry Halpin and Valentina Presutti. An ontology of resources: Solving the identity crisis. In *European Semantic Web Conference*, pages 521–534. Springer, 2009.
9. Afraz Jaffri, Hugh Glaser, and Ian Millard. URI disambiguation in the context of linked data. In *WWW Workshop on Linked Data on the Web, LDOW*, 2008.
10. Markus Nentwig, Michael Hartung, Axel-Cyrille Ngonga Ngomo, and Erhard Rahm. A survey of current link discovery frameworks. *Semantic Web*, 8(3):419–436, 2017.
11. Ngoc Thanh Nguyen. *Advanced Methods for Inconsistent Knowledge Management (Advanced Information and Knowledge Processing)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2007.