CHAPTER 1
INTRODUCTION

Since its adoption by Google in 2012, the term Knowledge Graph has rapidly evolved. Previously referring to a single project for semantically enhancing Google's search results [Singhal, 2012], this term currently refers to a wide range of graphs surging from academic research, community-driven efforts, and industrial projects, such as DBpedia[1], Wikidata[2], and the Facebook Social Graph[3]. Although Google have reaped the credits for its ever increasing popularity, the term knowledge graph has been around for years, making an appearance in Bakker's PhD dissertation [Bakker, 1987] as part of a Dutch project aiming at integrating and structuring scientific knowledge [Nurdiati and Hoede, 2008]. From a broad perspective, any graph-based representation of some knowledge in a machine-readable format, can be described as knowledge graph. However, many argue that knowledge graphs should fulfil certain requirements, necessary for enabling and enhancing various knowledge-based applications, such as semantic searches, intelligent chatbots, fraud detections, and recommendation systems. For instance, [Huang et al., 2017] mention size as a characteristic of knowledge graphs, while [Paulheim, 2017] requires the coverage of a major portion of domains, and [Färber et al., 2016] have restricted the use of this term to RDF[4] graphs. Adopting some of these proposed, more restrictive, definitions will affect the status of several existing knowledge graphs, since not all knowledge graphs are RDF graphs, or domain independent.

With the lack of a formal and standardized definition, a number of guiding principles have emerged for helping data publishers create high quality data and knowledge graphs. While some of the proposed principles, such as FAIR [Wilkinson et al., 2016], have provided a set of goals to ensure that published data are findable, accessible, interoperable, and reusable, independently of the technology used, other principles have acted as a set of methods and steps for publishing open and reusable data on the Web. The most known set of principles were laid out by Tim Berners-Lee in 2006, with the goal of encouraging people to use HTTP[5] IRIs[6] for naming things, and using W3C[7] standards for describing these IRIs (e.g. RDF(S) and OWL[8]), and linking them to other IRIs for providing context. This set of widely adopted principles, known as the Linked Data principles, refers to a set of best practices for publishing structured data on the Web so it can be easily interlinked and managed using semantic queries. The

[1] `https://wiki.dbpedia.org`
[2] `https://www.wikidata.org`
[3] `https://developers.facebook.com/docs/graph-api`
[4] Resource Description Framework
[5] Hypertext Transfer Protocol
[6] Internationalized Resource Identifiers
[7] World Wide Web Consortium
[8] Web Ontology Language

idea is by providing simple principles, for creating and publishing structured data, publishers can also enrich, access, and benefit from a larger decentralized knowledge graph, known as the Web of Data.

Despite the adoption of the Linked Data principles, achieving the FAIR goals still poses a number of significant practical and research challenges, particularly in terms of the interoperability and re-usability of the published data. Firstly, adopting standard knowledge representation languages for expressing, explicit and implicit, domain knowledge still poses particular challenges. Specifically, when dealing with complex domains such as medical and life sciences data, there is a need to express certain types of axioms and relations, that can not be intuitively expressed in even some of the most expressive standardized languages, such as OWL 2 DL. These limitations in the language prompts various research questions discussed in [Krisnadhi et al., 2015], and poses several challenges for modellers to express the necessary knowledge using current standards and best practices. In addition, and while adopting such standardized knowledge representation languages guarantees interoperability at a syntactic level, one of the important challenges consists in achieving interoperability at the semantic level [d'Aquin and Noy, 2012]. Semantic interoperability is the ability to meaningfully and accurately exchange and interpret information produced by different sources. Creating semantically interoperable knowledge graphs requires considerable efforts, and poses several practical challenges for modellers in finding, evaluating and reusing existing well-established models to describe their data.

Finally, achieving semantically interoperable knowledge graphs requires making links to other people's data. Such semantic interlinking is typically performed by asserting that two names (IRIs) denote the same real world entity. For this purpose, the Web Ontology Language OWL have introduced the `owl:sameAs` identity predicate. For instance, the triple $\langle President\_Barack\_Obama, \texttt{owl:sameAs}, 44th\_US\_president \rangle$ asserts that both names actually refer to the same person. Such identity statements indicate that every property asserted to one name will be also inferred to the other, allowing both names to be substituted in all contexts. While such inferences can be extremely useful in enabling and enhancing knowledge-based systems, incorrect use of identity can have wide-ranging effects in a global knowledge space like the Semantic Web. With studies dating back to the early Linked Data days showing that `owl:sameAs` is indeed misused in the Web [Jaffri et al., 2008, Ding et al., 2010a, Halpin et al., 2010], one can trace back their presence to several factors. Firstly, most `owl:sameAs` links are generated by heuristic entity resolution techniques, that employs practical strategies which are not guaranteed to be accurate. For instance, an algorithm matching books based on the similarity of their titles and authors is not always accurate, as two different editions of the same book can also share both these traits without being the same, since they do not share the same number of pages. In addition,

identity does not hold across all modal contexts, as things can be considered identical for some people in certain contexts, while being different in other contexts. For instance, drugs sharing the same chemical structure, but produced by different companies, are considered identical in a scientific context, but are different in a commercial one.

Since suitable alternatives to `owl:sameAs` have yet to exist, or are rarely used in practice, a given Linked Data application is forced to make a choice with respect to each `owl:sameAs` assertion it encounters. This problem of incorrect usage of identity is not specific to the Semantic Web, and is present in all Knowledge Representation systems [Grant and Subrahmanian, 1995, Nguyen, 2007]. However, the problem is specifically alerting in the Semantic Web 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 recognized, and has been referred to as the "Identity Crisis" [Bouquet et al., 2007], and the "sameAs problem" [Halpin et al., 2010]. As such, a proper approach towards the handling of identity links is required in order to make the Semantic Web succeed as an integrated knowledge space.

1.1 Objectives & Contributions

Identity management in knowledge graphs is the main objective of this thesis. Despite its ambitious title, this thesis is a modest attempt to address one particular issue of the identity problem: the excessive and incorrect use 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 [Ferrara et al., 2013] and frameworks [Nentwig et al., 2017] 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 [Halpin, 2010]. 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** [Beek et al., 2018]. In order to uncover different aspects of the use of identity in the Semantic Web, and to facilitate access to a large number of identity statements, we propose `sameas.cc`: a web service and a dataset containing the largest number of identity statements that has been gathered from the LOD Cloud to date. This service provides public access (query and download) to over 558 million distinct `owl:sameAs` statements extracted from the LOD Cloud. It also provides access to these links' equivalence closure, and the resulting identity sets. For this, we propose an efficient approach for calculating and storing the equivalence closure, that exploits the `owl:sameAs` transitive semantics.

Both explicit identity statements, and their equivalence closure are accessible at `http://sameas.cc`.

- **Approach for detecting erroneous identity links** [Raad et al., 2018a, Raad et al., 2018b]. With many studies already showing that identity links are incorrectly used in the Semantic Web, there is an ever increasing need to detect these links to ensure the quality of knowledge graphs. For this, we propose an approach for automatically detecting potentially erroneous identity links, by making use of the `owl:sameAs` network topology, and more specifically the network's community structure. Based on the detected communities, an error degree is calculated for each identity link which is subsequently used to rank identity links, allowing potentially erroneous links to be identified, 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 evaluation is run on the `sameas.cc` dataset. The implementation of this approach is available at `https://github.com/raadjoe/LOD-Community-Detection`.
- **A contextual identity relation** [Raad et al., 2017a, Raad et al., 2017b]. 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. For instance, in some applications, the fact that drugs share the same chemical structure is sufficient to consider them as equivalent, while in other applications it is also necessary that these drugs share the same name. Unfortunately, modelling the specific contexts in which an identity relation holds is cumbersome and, due to arbitrary reuse, and the Open World Assumption, it is impossible to anticipate all contexts in which an entity will be used. For this, we define a new contextual identity relation. This relation expresses an identity between two class instances, that is valid in a context defined regarding a domain ontology. For automatically generating these contextualized identity assertions, we propose an algorithm named `DECIDE`. This algorithm detects the most specific contexts in which a couple of instances are identical. In addition, and since not all contexts are relevant (e.g. a context considering a value without its unit of measure), this algorithm can be guided by different sets of semantic constraints provided by experts for enhancing the detected contexts. The implementation of this approach is available at `https://github.com/raadjoe/DECIDE_v2`.
- **Contextually linked knowledge graphs for life sciences** [Ibanescu et al., 2016, Raad et al., 2018c]. Cases in which objects can not be declared the same are quite common in scientific data, 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 serious barrier for knowledge-based systems 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 classic problems of the heterogeneity of the formats in which scientific data are published, and the terminological variations encountered across the multiple scientific datasets remain serious barriers in fully exploiting the large amount of data produced everyday. As a way for limiting these syntactic, semantic and identity problems, we introduce a new knowledge graph for life sciences. This graph is constructed in a mutual effort with domain experts of 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 contextual identity links. In addition, we exploit the millions of detected contextual identity links in this 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. This knowledge graph can be queried and downloaded at `http://sonorus.agroparistech.fr:7200`.

1.2 Thesis Outline

This classic identity problem, recently amplified in the context of Linked Data, has led to several analysis, discussions, and proposals for limiting its effects. Chapter 2 gives an overview on the existing solutions in the Semantic Web, and reflects on the current state of this problem. Chapter 3 presents our first contribution to this problem, by introducing the `sameas.cc` dataset and web service, which we deploy for performing several analyses on the use of identity in the LOD Cloud. Chapter 4 presents our approach of detecting erroneous identity links using network metrics, and the experiments conducted on a large subset of the LOD Cloud. Chapter 5 introduces our new contextual identity relation, and presents our approach for automatically detecting these links in an RDF knowledge graph. Chapter 6 presents a new knowledge graph for life sciences, and describes the exploitation of the detected contextual identity links for discovering certain rules, and predicting observational measures. Chapter 7 discusses the results of the research presented in this thesis, as well as its limitations, lessons learned during the process of conducting it, and some lines for future work.