le cnam

Conservatoire National des Arts et Métiers

Habilitation à Diriger des Recherches

(Spécialité Informatique)

Ecole doctorale SMI - Sciences des Métiers de l'Ingénieur

Enrichment of RDF Knowledge Graphs with Contextual Identity Links and Fuzzy Temporal Data

Présentée par : Fayçal Hamdi

Rapporteurs :

Bernd AMANN, Professeur des Universités, Sorbonne Université
Nathalie AUSSENAC-GILLES, Directrice de Recherche CNRS, IRIT Toulouse
Farid MEZIANE, Professeur, Université de Derby, Royaume-Uni

Examinateurs :

Kamel BARKAOUI, Professeur des Universités, CNAM
Raja CHIKY, Directrice de Recherche, ISEP
Myriam LAMOLLE, Professeure des Universités, Université Paris 8
Elisabeth METAIS, Professeure des Universités, CNAM (Marraine)

Soutenue le : 5 novembre 2020

Abstract

This HDR thesis summarizes approaches and tools developed during the last eight years of my research activities as well as future directions of my research projects.

In the context of the Semantic Web, the widespread use of semantic web technologies leads to the publication, on the Web of Data, of billions of data in the form of a graph, called RDF knowledge graphs. However, for many users or automated agents, enrich and take advantage of knowledge graphs may not be as obvious as it appears. Indeed, enriching the knowledge graph with data that does not take into account the context or that is not well represented can lead to a deterioration in the quality of the data, and thus, in the quality of inferred conclusions or decisions. Among the different data that could enrich a knowledge graph, there are identity links and temporal data. First, concerning identity links, the main issue is related to the different parameters used to interlink entities and to the type of links. Hence, in some domains, using basic distance measures could not be sufficient to establish links between entities. We verified this fact in the context of geographic data that are present in many existing knowledge graphs. In this domain, when the geographic characteristic of entities are captured in an ambiguous way, basic spatial distance-based matching algorithms may produce erroneous links. To deal with this issue, we suggested formalizing and acquiring the knowledge about the spatial references, namely their positional accuracy, their geometric modeling, their level of detail, and the vagueness of the spatial entities they represent. We then proposed an interlinking approach that dynamically adapts the way spatial references are compared, based on this knowledge. Furthermore, the type of the link established between two entities is very important. Numerous interconnections between knowledge graphs, published on the Web of Data, use the *owl:sameAs* property which supposes that linked entities must be identical in all possible and imaginable contexts. Since identity is context-dependent, all property-value pairs of the first entity could be wrongly transferred (propagated) to the other entity. Thus, we proposed an approach, based on sentence embedding, to semi-automatically find a set of properties, for a given identity context, that can be propagated between contextually identical entities. Second, concerning temporal data, the problem that we addressed concerns the enrichment of knowledge graphs with imprecise data (e.g., "late 1970s"), which requires the use of fuzzy logic. We proposed, using this logic, an ontology-based approach for representing and reasoning about precise and imprecise temporal data. Both quantitative (i.e., time intervals and points) and qualitative (i.e., relations between time intervals, relations between a time interval and a time point, etc.) temporal data were considered. Finally, as inferred conclusions or decisions depend on the quality of data, we proposed to evaluate the quality of knowledge graphs regarding two important quality dimensions that are completeness and conciseness. For the completeness, we proposed a mining-based approach to derive a suitable schema (i.e., a set of properties) that will be used in the calculation of this dimension. We implemented a prototype called "LOD-CM" to illustrate the process of deriving a conceptual schema of a given knowledge graph based on the user's requirements. Concerning the conciseness, its calculation is based on discovering equivalent predicates. Thus, to find these predicates, we proposed an approach based, in addition to a statistical analysis, on a deep semantic analysis of data and on learning algorithms.

All these works were developed in the context of PhD and Master theses as well as collaboration with many researchers in the context of different research projects.

Keywords : Semantic Web, Knowledge Graphs, Linked data, Contextual Identity, Ontology, Temporal Representation, Temporal Reasoning , Fuzzy Ontology, Sentence Embedding, Completeness, Conceptual Schema Mining, Conciseness.

Acknowledgements

First of all, I would like to thank my HDR reviewers Prof. Bernd Amann, Dr. Nathalie Aussenac-Gilles, and Prof. Farid Meziane who accepted to review and evaluate this dissertation.

Many thanks to Prof. Kamel Barkaoui, Dr. Raja Chiky, and Prof. Myriam Lamolle for being a part of my examining committee.

I am deeply grateful to Prof. Elisabeth Métais for supporting me in the realization of this HDR.

My sincere thanks go to my colleagues in the ISID team, the CEDRIC Laboratory, and the computer science department (EPN5) of the Conservatoire National des Arts et Métiers.

Many thanks to my PhD students for several years of constructive and enriching discussions and to my colleagues with whom I collaborated on various projects.

Last but not least, I am thankful to my wonderful family for their unlimited support and encouragement.

Table des matières

1.1 General Context

In the context of the Semantic Web, the widespread use of semantic web technologies leads to the publication, on the Web of Data, of billions of data in the form of a graph, called Linked Data or most recently "Knowledge Graphs" (KGs). This new term was initiated in 2012 by Google [1]. While many definitions were proposed, none of them has been unanimously adopted. [EW16] propose a survey of the most common definitions. All these definitions shared the following statements : *(i)* it must be a graph *(ii)* that describe entities from the real world or not, and *(iii)* relationships between those entities. It is the case of datasets published on the Linked Open Data cloud [2] (LOD) following the Tim Berners-Lee's principles and its vision of the Semantic Web. These datasets are called today "RDF Knowledge Graphs".

For many users or automated agents, enrich and take advantage of Knowledge Graphs may not be as obvious as it appears. In fact, enriching the Knowledge Graph with data that does not take into account the context or that is not well represented can lead to a deterioration in the quality of the data, and thus, in the quality of inferred conclusions or decisions. Among the different data that could enrich a Knowledge Graph, there are identity links and temporal data.

Creating identity links is related to the identification of links between resources which represent the same real-world entity, or are related to each other by some kind of relationship. The idea consists of comparing the values of similar properties used by resources from heterogeneous data sources for describing real-world entities in order to estimate the degree of similarity between these resources. The higher the similarity score is between two resources, the more they are likely to represent the same real-world entity [FNS13].

In some domains, using basic distance measures to create identity links could not be sufficient to establish links between entities. It is the case of geographic data where published resources are associated, via geometries (spatial references), to a location in the geographic space. These geometries may be direct, such as geographic coordinates or geometric primitives (points, linestrings or polygons), or indirect such as postal addresses or names of administrative units. Like any other property, spatial references can be used to evaluate the similarity of resources in a data matching process. In the field of geographic data matching, many measures have been proposed to evaluate the similarity of geometries. They are progressively implemented in data linking tools. However, they have been designed for traditional geographic databases matching and may not provide good results when directly reused for georeferenced resources of the Web. The open nature of Web of data sources, mainly produced by crowdsourcing, raises new challenges for geometric similarity evaluation due to the heterogeneity of the geometric quality within and between data sources. Indeed spatial data on the Web may have various origins : they may be extracted or transformed from geographic databases provided by traditional data producers such as national mapping agencies (e.g. Ordnance Survey data, geo.linkeddata.es), but they may also be a fusion of many data sources as they can be produced by crowdsourcing (e.g. DBpedia [3], Geonames [4]). In such cases, inside the same data source, spatial references may have been captured differently which leads to what we call "internal heterogeneity". In this case, defining and applying the same interconnection process, with the same parameters for all the processed resources, can then become extremely complex, or even lead to erroneous links. To tackle this problem, a possible solution is to consider the heterogeneity of geometries in the interlinking process.

Furthermore, the type of the link established between two entities is very important. One way to link two entities within the same graph or to link several Knowledge Graphs is to use the OWL property *owl:sameAs* which states that two entities represent the same two real-world objects. This kind of link helps promoting the reuse of these entities and their descriptions and allows to discover new information.

1. https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html - Accessed : 2019-10-31
2. http://lod-cloud.net
3. http://www.dbpedia.org/
4. http://www.geonames.org/

Indeed, a property-value pair describing an entity can be reused by any other identical entity. This reuse is one of the strong points of this property since *owl:sameAs* makes it possible to increase the completeness of an entity, i.e., to increase the knowledge that one has about it. However, the semantics of *owl:sameAs* is strict, since, to be identical, two entities must have the same property-value pairs in all possible and imaginable contexts. This conception of identity has been debated many times [HHM+10, DSFM10] and brings some challenges to the Semantic Web framework. Indeed, in many cases this property is problematic, since the identity relationship is often context-dependent. As the *owl:sameAs* property is, by definition, transitive, it is possible to have several identical entities linked by *owl:sameAs* and forming a chain of identical entities where any property-value pair can be used on any entity of this chain. This mechanism is called "property propagation". As a result, a link that may sometimes be true and sometimes false risks spreading false information. The propagation of false data along a chain of *owl:sameAs* properties decreases the quality of the data and could have dramatic consequences depending on data usage. Thus, it is important to take into account the context (a detailed definition is given in Chapter 3) when creating identity links.

Another category of data that is widely represented in Knowledge Graphs is temporal data. Enriching KGs with this kind of data and reasoning over them could be a very complex process especially when the information to describe is imprecise. For instance, the sentence "*Alexandre was married to Nicole by 1981 to late 90*" conveys two kinds of imprecision : *(i)* the information "*by 1981*" is imprecise in the sense that it could mean approximately from 1980 to 1982, and *(i)* "*late 90*" is imprecise in the sense that it could mean, with an increasingly possibility, from 1995 to 2000. When an event is characterized by a gradual beginning and/or ending, it is usual to represent the corresponding time span as an imprecise time interval. To consider this graduation in the representation and reasoning on this kind of data, a possible solution is to use fuzzy logic.

Finally, as inferred conclusions or decisions depend on the quality of data published on Knowledge Graphs, it is important to provide methods and tools to assess this quality. Information quality has been extensively studied in the context of relational databases [WS96, LSKW02]. However, in the context of Web of Data, this is an emerging issue and a real challenge for future research. Indeed, Linked Open Data through HTTP URI's offer an infrastructure for publishing and navigating through data without ensuring the quality of its usage [BBDR+13]. As a consequence, there is a real need for developing suitable methods and techniques for evaluating and helping ensure web data quality.

This HDR thesis outlines my contributions to the enrichment of Knowledge Graph with contextual identity links and temporal information as well as the quality assessment of data published in KGs. The presented results were obtained from 2012 to early 2020, as an Associate professor at Cnam Paris (Cedric Laboratory, EA 4629, ISID Team) through collaborations with colleagues, five PhD students and more than twenty master students (cf. Curriculum Vitae in Annex A).

1.2 Contributions

After my work on ontology alignment carried out during my doctoral thesis [Ham11], I focused my research on issues related to the contextual identity, temporal data, and quality assessment. All the approaches I present in this manuscript have led to the development of tools that have been used to conduct experiments on real-world large-scale Knowledge Graphs. Hence, my contributions to the Knowledge Graphs enrichment are the following :

— An adaptive interlinking approach in the geographic domain based on the formalization of heterogeneities in the representation of the geometries.

— A context-related representation of identity links based on an approach that identifies properties that can be propagated in a given context.

— An approach to represent and reason about imprecise temporal data.

— Approaches to assess the completeness and conciseness of Knowledge Graphs.

Domain (geographic domain) adaptation of the interlinking process. To adapt the interlinking process

to the geographic context, we propose to formalize knowledge about the characteristics of spatial references (geometries) to identify potential heterogeneities. We thus propose a vocabulary allowing to associate to any spatial reference used to locate a resource, metadata that describes the absolute positional accuracy of geometries, the geometry capture rules (geometric modeling), the vagueness of the spatial characteristics of the geographic entities represented by the geometries, and the level of detail of the data sources. Having knowledge of the characteristics of spatial references, we can automatically deduce the level of heterogeneity that two spatial references are likely to present and thus adapt the interlinking approach. We therefore propose an approach for interconnecting georeferenced resources using this knowledge. This approach implements a rule-based reasoning on this knowledge to dynamically adapt the parameterization of the spatial reference comparison during the interconnection process. Experiments have been carried out on real-world data to validate our approach.

Context-related identity links. To take into account the context in identity links, we proposed an approach that finds semi-automatically the properties that can be propagated between identical entities in a given context. We use the definition of an identity context proposed by [IHvH+17] since it provides a framework for property propagation for a set of indiscernible properties. Our approach uses a sentence embedding technique where a vector represents each property having a long description in natural language. It is thus possible to compute a vector representing a set of indiscernible properties, on the one hand, and, on the other hand, to compute the distance between this vector and those of the properties that are candidates for propagation. We postulate that the closer the description of a property is to the descriptions of the indiscernible set, the more likely it is propagable. To find the contexts of an entity, we propose an algorithm that computes the lattice that represents the set of identity contexts of this entity. To evaluate our approach, a quantitative and qualitative experiments have been conducted.

Representation and reasoning about imprecise temporal data. To address the problem of enrichment of KGs with temporal data, we propose two approaches. The first one involves only crisp standards and tools. To represent imprecise time intervals in OWL 2, we extend the so called 4D-fluents model [WF06] which is a formalism to model crisp quantitative temporal information and the evolution of temporal concepts in OWL. This model is extended in two ways : *(i)* with new crisp components for modeling imprecise time intervals, and *(ii)* with qualitative temporal expressions representing crisp relations between imprecise temporal intervals. To reason on imprecise time intervals, we extend the Allen's interval algebra [All83] which is the most used and known theory for reasoning about crisp time intervals. We generalize Allen's relationships to handle imprecise time intervals with a crisp view. The resulting crisp temporal interval relations are inferred from the introduced imprecise time intervals using a set of SWRL rules [HPSB+04], in OWL 2. The second approach is based on fuzzy sets theory and fuzzy tools. It is based on Fuzzy-OWL 2 [BS11] which is an extension of OWL 2 that deals with fuzzy information. To represent imprecise time intervals in Fuzzy-OWL 2, we extend the 4D-fluents model in two ways : *(i)* with new fuzzy components to be able to model imprecise time intervals, *(ii)* with qualitative temporal expressions representing fuzzy relations between imprecise temporal intervals. To reason on imprecise time intervals, we extend Allen's work to compare imprecise time intervals in a fuzzy gradual personalized way. These two approaches have been implemented and tested in Fuzzy-OWL 2 using an ontology editor and a fuzzy reasoner FuzzyDL.

The KGs quality assessment. We studied two dimensions of Knowledge Graphs quality that are completeness and conciseness. This choice is motivated by the fact that these two dimensions affect other data quality dimensions, such as accuracy, and impacts directly users' queries. Concerning the completeness, as it often relies on gold standards and/or a reference data schema that are neither always available nor realistic from a practical point of view, we propose an assessment approach that uses mining algorithms to extract a schema from data values. The proposed approach is a two-step process : first, mining from a dataset a schema that reflects the actual representation of data, which, in the second step, is used for completeness evaluation. For assessing the conciseness of a KG throw the identification of equivalent predicates, we propose an approach that consists of three sequential phases : a statistical analysis to obtain an initial set of synonym predicates, and a semantic and an NLP-based analysis to improve the precision of the identification of synonyms. Prototypes have been implemented to evaluate our assessment approaches.

1.3 Manuscript Outlines

I chose to present **some of my research work I did in the past eight years, by focusing only on those that are related to Knowledge Graph enrichment and quality assessment**. The manuscript is structured as follows :

Chapter 2. I present in section 2.1 related works about geometry similarity evaluation on the Web of data. In section 2.2, I present a vocabulary to describe geometry characteristics that must be taken into account for geometry comparisons and in section 2.3 I detail our adaptive geometry similarity evaluation approach. Section 2.4 details the experiments that we carried out to validate our approach.

Chapter 3. After introducing the objective of our propagation approach and highlighting the challenges through an illustrative example, I present the related work in section 3.1. Then in section 3.3, I present our approach and in section 3.4, the quantitative and qualitative experiments we have conducted.

Chapter 4. In section 4.1, I present some preliminary concepts and related work in the field of temporal information representation in OWL and reasoning on time intervals. In section 4.2, I introduce our crisp-based approach for representing and reasoning on imprecise time intervals. In section 4.3, I present our fuzzy-based approach for representing and reasoning on imprecise time intervals.

Chapter 5. I summarize in section 5.1 related literature on completeness assessment for KGS, detail our mining-based approach for RDF data conceptual modeling, and present two use cases for the LOD-CM, a web tool that we developed to validate the approach. In section 5.2, I present related work about the conciseness dimension, describe a motivating example, explain our proposed approach that consists of three sequential phases, and finally present two experiments performed on real-world datasets to evaluate our proposed approach that aims to assess conciseness of Linked Data.

Chapter 6. Summarizes my contributions and gives some directions for future work.

Enriching KGs with Domain Identity links

This chapter is based on work realized in collaboration with The French national mapping agency (IGN) in the context of the PhD thesis of Abdelfettah Feliachi that I co-supervised with Dr. Nathalie Abadie and Dr. Bénédicte Bucher (from IGN France). The result of this work was published in : [FAH17a], [FAH17b], [FAH14], [FAHA13]

Web-based data resources are often equipped with spatial references such as place names, addresses or geographic coordinates. Like any other property associated with Web data resources, these spatial references can be used to compare resources and enrich Knowledge Graphs with identity links. Current approaches to detecting matching relationships between spatially referenced data are mainly based on the assumption that resources described by geographically close spatial references are more likely to represent the same real-world geographic entity. Thus, resources are compared using calculated distance measurements between their spatial references.

In contrast to spatial datasets from traditional data producers such as national mapping agencies, datasets published on the Web are mostly derived from participatory data sources or are the result of the aggregation of several datasets. Thus, the spatial references they contain are not necessarily all produced in the same way : their locational accuracy, the characteristic feature of the shape of the geographic feature taken as an input marker, or the type of geographic feature represented may vary from one spatial reference to another. In this case, defining and applying the same interconnection process, with the same parameters for all the resources processed, can then become extremely complex, or even lead to erroneous links. In this work, we follow the intuition that improving the spatial data matching results requires one to adapt each pair of geometries similarity evaluation to the characteristics of the tested geometries. We thus suggest to use metadata, describing the causes of heterogeneities, to automatically adapt the matching process. These metadata include : the positional accuracy, the geometric modeling, the level of detail, and the vagueness of spatial entities. The main contributions of this work are :

(1) a new vocabulary that describes the causes of geometric heterogeneities that should be taken into account in the setting of a spatial data matching process. We called this vocabulary "The XY Semantics Ontology".

(2) an approach based on supervised machine learning to extract the causes of heterogeneities from the data to be matched when information about their geometries capture process is not provided.

(3) a data matching approach that uses knowledge about the causes of geometric heterogeneities to dynamically adapt the parameters used to compare the geometries. This approach is experimentally evaluated on two datasets about historical monuments and we compare its results with results obtained with a fixed-parameters approach.

The remainder of this chapter is organized as follows : section 2.1 presents related works about geometry similarity evaluation on the Web of data. In section 2.2, we present a vocabulary to describe geometry characteristics that must be taken into account for geometry comparisons and in section 2.3 we detail our adaptive geometry similarity evaluation approach. Section 2.4 details the experiments that we carried out to validate our approach.

2.1 State of the Art

Previous works in the field of geographic data matching have focused on approaches for evaluating the similarity of geographic features, mainly based on the evaluation of their geometries similarity. As two geographic feature located far from each other in space are not likely to represent the same real world phenomenon, geometry similarity prevails indeed over any other properties similarity to decide whether two geographic features should be matched or not [ALRG10]. In this section, we present the approaches proposed in the field of geographic data matching for geometric similarity evaluation in order to adapt

them for georeferenced resources linking.

2.1.1 Geometry similarity measures for geographic data matching

Geographic databases are created through a process of abstraction of real world phenomena. Geometries are used to provide a quantitative description of spatial characteristics of real world entities, such as their dimension, position, size, shape, and orientation [PRL+03]. Due to the quality of raw data sources and the discrete nature of the geometrical primitives in spatial databases, the spatial characteristics of real world entities can be captured only in a simplified way. This has an impact on the quality and the data capture rules of the resulting geometries [Gir12].

Geometry similarity measures designed for geographic databases matching are all based on one or more geometry similarity function(s), that evaluate(s) geometry similarity with respect to some particular descriptor [JRK11]. These functions are chosen depending on the types of the geometries and the criteria with respect to which they are compared : distance functions based on euclidean, orthodromic or elliptic curve spatial distances, like the min distance function, deal with location of points sets [SN15] ; boolean functions based on the inclusion of the evaluated geometry in some buffer built around a given geometry are used in [WF99], [HB10] and [VBSC12] for comparing geometry locations ; the surface distance is used by [BHA99] to compute the similarity between polygons with regards to their location and the area of their overlapping surfaces ; [CG97] and [ACH+91] propose two measures for comparing polygon shapes, respectively based on distances and angles values ; the Hausdorff and Fréchet distances, which deal with both location and shape of linestrings, are widely used for linestring and polygon similarity evaluation (see for example [SH11], [Cos14]) ; [RMR15] uses a function for comparing linestrings orientation and [WCZ+15] functions for comparing polygons orientation, area and length ; [Cos14], [RMR15] also use functions for comparing geometries neighborhood (i.e. their topologically related or spatially close geometries).

All geometry similarity measures use at least one location-based similarity function. This function can be combined with functions based on other descriptors. To that end, they are standardized to values between 0 and 1 by means of various normalization functions and (eventually weighted) aggregation methods are used to compute an overall standardized geometry similarity value [JRK11].

Parameters such as buffer size or normalization function thresholds are used to define to what extent differences between geometries with regards to some given descriptor are considered acceptable. Setting such parameters is usually assigned to experts, who define their values based on their knowledge about the databases to be matched and the functions behaviors. Parameters related to the evaluation of location-based geometry similarity are the most intuitive. Most of the time, they represent the maximum acceptable spatial distance between two geometries for them to be considered as potentially representing the same real world entity ; above this value, geometries are considered too far each other to represent the same thing. This parameter is thus closely related to the absolute positional planimetric accuracy of the databases, defined by the ISO 19157 standard as the ”closeness of reported coordinate values to values accepted as or being true”. It may also be affected by geometry capture rules or geographic feature boundary vagueness. For example, a postal address represented by a point might be captured in various ways : within the extent of the building located by the address, at the entrance of the building, on the centerline of the street in front of the building, etc. [RMR15] details what information is needed to configure confidence functions used for geometry similarity evaluation based on location, neighborhood and orientation.

Most of geometry similarity measures designed for complex geometries such as linestrings and polygons combine several geometric similarity functions [RMR15], [WCZ+15]. This is usually done to overcome some data integration conflicts due to differences in the levels of detail of the datasets, i.e. the degree of geometric and semantic abstraction used for representing real world entities in these geographic datasets [Sar07]. Geographic databases integration conflicts caused by differences of level of detail have been thoroughly described by [DPS98]. In [Vol06] and [MD08], the road networks to be matched have different levels of granularity, i.e. the road segments are more detailed in one of the databases than in the other. This conflict is solved by the similarity functions based on the spatial neighborhood of road edges and nodes.

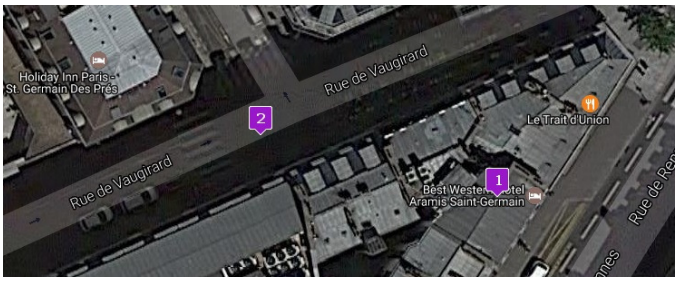

FIGURE 2.1 : Internal geometric heterogeneity in Geonames.

Some approaches also apply geometry transformation operations before computing geometry similarity in order to reduce the gap between the geometries to be compared. For example, [WF99] performs conflation to lower the location differences between the road segments to be matched due to each database positional planimetric accuracy. [YLZ14] uses generalisation algorithms to harmonise the levels of detail of the compared databases.

Many approaches have thus been proposed to compute geometry similarity for geographic databases matching. Some of them are progressively introduced for georeferenced resource linking.

2.1.2 The heterogeneity of geometries on the Web of data

Unlike geometries of geographic databases, geometries used on the Web of data are not the main piece of information of the resources they describe. They are not necessarily available in the description of the resources. Besides, spatial data on the Web may have various origins : they may be extracted or transformed from geographic databases provided by traditional data producers such as national mapping agencies (e.g. Ordnance Survey data, geo.linkeddata.es), but they may also be a fusion of many data sources as they can be produced by crowdsourcing (e.g. DBpedia [1], Geonames [2]). In such cases, inside a same data source, spatial references may have been captured differently which leads to what we call "internal heterogeneity".

Standardization efforts, specifically the recommendation of the OGC/W3C working group on the best practices when publishing spatial data [3] tend to solve syntactic heterogeneity of geometries on the Web of data. However, crowdsourced geometries may have been produced with different capture rules and at different levels of details. Fig. 2.1 shows an example of internal heterogeneity of geometries. In Geonames, hotels are located by points. One was captured on the building while the other was captured on the roadway. As Geonames is an open voluntary data source, this discrepancy could be due to some difference of positional accuracy or geometry capture choices.

Classical geometry similarity measures presented in 2.1.1 are designed to compare homogeneous geometry sets, each set being produced at the same level of detail, with the same positional planimetric accuracy and the same capture rules. They may thus be inadequate for geometries used to georeferenced linked data.

2.1.3 Geometry similarity functions selection, combination and tuning

Choosing automatically the adequate setting of a matching process has been the subject of many works in both ontology matching and data linking fields, and still poses challenges [SE13]. Approaches such as [FES14] and [NUMDR08] propose to take advantage of the alignment of the vocabularies that structure the data to select the properties that should be compared and the distance measures used for that purpose. Other approaches addressed the question of tuning the parameters of the matching process such as the approach proposed in [SNA+15] to automatically compute the comparison criteria weights. The challenge of tuning the ontology matching process has been addressed by various approaches. [DBC08] propose an approach for choosing the comparison criteria and computing a decision tree to aggregate them. [HHT07] propose a classification of matching algorithms and use a set of decision rules to assess to each ontology context an adequate matching algorithm. [MJ08] propose also an approach based on decision rules to tune

1. http://www.dbpedia.org/
2. http://www.geonames.org/
3. https://www.w3.org/TR/sdw-bp/

the matching process using the metadata describing the ontologies and those describing the matching algorithms

The self-tuning of the matching process provides an adaptation of its settings to the context of matching while reducing the intervention of the expert. In fact, techniques such as using decision rules provide a materialization of the experts' knowledge about the best setting of a matching process according to different contexts. In this work, our intuition is to use decision rules to automatically select and tune the adequate similarity measures between geometries while taking into account the heterogeneities that may exist between every pair of geometries.

2.2 The XY Semantics Ontology

We have seen in the previous section that geometries used for representing real world geographic features may be produced by different capture processes and may therefore be different from one data source to another. In addition, human and material input errors and the collaborative open nature of some data sources on the Web may accentuate geometric heterogeneities within a single dataset. A geometric level of detail and well-defined data specifications allow us to understand the meaning of each geometry : what it represents, how it was captured, how it is modeled, how accurate it is, and so on. In other words, what is the semantics carried by this geometry ? The heterogeneities between the geometries are therefore nothing but differences in their semantics. We thus define **"the semantics of the XY"** as *"the set of geometry characteristics related to the geometric level of detail and to the capture process"*. Naming our ontology as so is motivated by the definition of geographic data semantics given by [KK07] as the relationship between the data and the real world phenomenon they represent.

From the heterogeneities faced in the geographic databases matching approaches, to the challenges faced in the context of the Web of data presented in section 2.1, we have identified the following characteristics that are more likely to affect the setting of a spatial data matching process :

— the absolute positional accuracy of geometries,

— the geometry capture rules (geometric modeling),

— the vagueness of the spatial characteristics of the geographic entities represented by the geometries,

— the level of detail of the data sources.

2.2.1 Vocabulary description

We propose an ontology [4], called XY semantics, that describes these characteristics, and thus, enables using them as knowledge through an interconnection process. We have chosen only these four characteristics since we assume that, although they are the most important for understanding heterogeneities, the only way to take advantage of them is to make them explicit. Indeed, other geometric features such as orientation, elongation, area, etc. are implicitly present in geometry, and therefore they are not difficult to extract on the fly.

The XY semantics ontology is based on the ISO standards on geographic databases metadata ISO 19115 and geographic data quality ISO 19157. It also includes works related to spatial entities vagueness [Sch10] and geometry capture rules [Aba12]. Fig. 2.2 shows an excerpt of XY semantics ontology.

The XY semantics ontology enables to associate to each geometry elements describing each of the four characteristics of geometry semantics listed above. As an example, the positional planimetric accuracy is described by a method and an evaluation result. The evaluation method specifies whether the evaluation result is derived from another quality element or assessed from the data. This latter is the most often employed. A data-based evaluation can be carried out by looking to a sample of the data or to their genealogy. For this reason we used the *Entity* class from the PROV-O [5] ontology. The evaluation result is described by its numerical value as well as its unit of measure defined from the QUDT [6] ontology.

4. http://data.ign.fr/def/xysemantics
5. `https://www.w3.org/TR/prov-o/`
6. `http://qudt.org/schema/qudt`

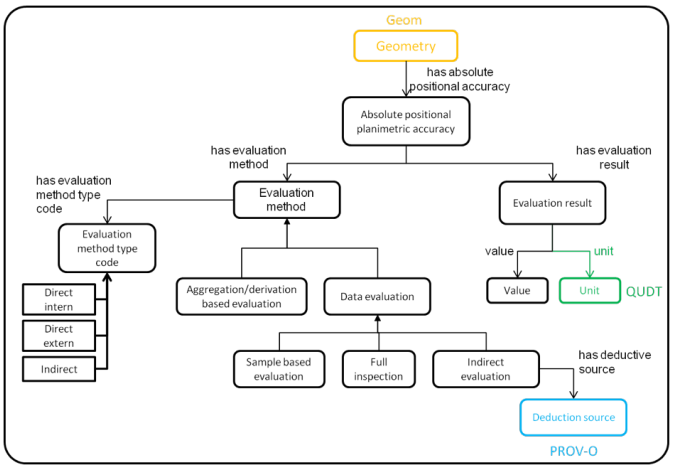

FIGURE 2.2 : Excerpt of the XY semantics ontology describing the planimetric accuracy of geometries.

Populating the XY semantics ontology is a task that can be very complicated. In the case of geographic data produced by mapping agencies, descriptive records and metadata, including geometric metadata, are often provided with datasets. Moreover, these datasets guarantee an internal homogeneity of the geometrical characteristics : for the same class, geometries often have often the same geometrical modeling, the same planimetric accuracy, etc. Even when the geometries of the same class have different characteristics, additional indications are often provided to explain the different situations (e.g. the different possible planimetric accuracy of addresses). In contrast, in the context of the Web of data, metadata about the methods used for georeferencing or its quality are rarely provided. In addition, datasets built collaboratively do not necessarily provide guidance on how to represent spatial references, and even if they do, they do not necessarily guarantee that contributors comply with these guidelines.

We describe in the following how to populate the XY semantics ontology in the presence and in the absence of geometric metadata.

2.2.2 Populating the ontology when geometric metadata are provided in datasets

The metadata of the geographical data are often provided in descriptive files. Those provided with the authoritative data of mapping agencies make populating our ontology much easier. According to the metadata of the IGN [7] address database, address points are captured in various locations : at the address sign, at the entrance of the building, 4.5m from the axis of the street (by projections from centroids of plots, by interpolations or arbitrarily), in an addressing area or in the center of the city. Moreover, the different planimetric accuracy values of the geometries are provided. These metadata can be easily translated into RDF data structures according to our XY semantics ontology and associated them with each geometry through SPARQL data insertion queries.

2.2.3 Populating the ontology when geometric metadata are not provided

Identifying geometric characteristics when they are not described in metadata is a laborious task if performed manually for each geometry in a data source. Thus, populating the XY semantics ontology becomes a complicated task. To deal with this issue, we propose a two-steps approach that automatically identifies the geometry capture rules. First, for each resource within the same dataset it finds which characteristic element of its form was chosen, when the coordinates used to locate it were entered. Identifying this characteristic allows then the evaluation of the planimetric accuracy of the spatial references. This latter is carried out by adapting the direct estimation methods for the absolute planimetric accuracy of geographical data [ISO13] to the collaborative data.

We propose to use a ”reverse-engineering” mode to identify the different geometric modelings of the spatial references. We start from the main assumption that a geometry results from an intentional choice

7. The French national mapping agency

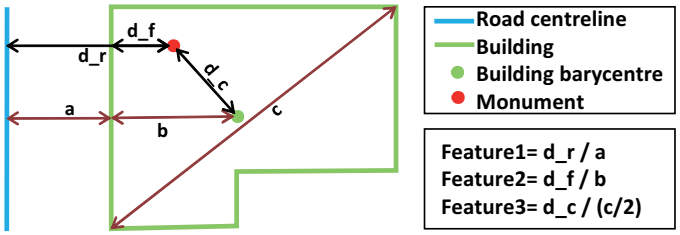

FIGURE 2.3 : Learning features for geometry capture rules

of geometric modeling by the contributor. We propose to formulate the different hypotheses on the choices made by the contributors when entering geometries of resources. These hypothetical choices may be determined by visually comparing the spatial references of resources to the geometries used to represent semantically close or equivalent geographic types within a reference geographic dataset. This empirical visual analysis allows mainly to identify the various hypothetical patterns (trends) of the geometric modeling choices that emerge from the data.

The next step is to associate each spatial reference to one of the identified geometric modeling patterns. We formalize our problem as follows : we have a geometry population G and a set of geometric modeling classes $\{C_1,...,C_n\}$. We must therefore define a set of relevant descriptors D and select a set of learning $S \times \{C_1,...,C_n\}$ (with $S \subset G$) in order to define the classification function C. Relevant descriptors D must be descriptive indicators whose values combination allow to discriminate the different classes. To define them, we propose to analyze the geometries of resources in comparison with geometries which represent geographical entities of semantically similar or equivalent types within a reference geographic dataset. The descriptors can therefore be a distance or a relationship between the analyzed geometries and the characteristic elements of the shape of the geographical features represented by reference geometries. We can for example consider the distance between each analyzed geometry and the closest linestring used for representing a road centerline in the reference dataset. The selection of a learning set consists in finding for each class of geometric modeling a representative sample of easily recognizable geometries in the analyzed dataset. Then, we apply a learning algorithm to assign each geometry to a geometric modeling class.

In order to evaluate our approach, we applied it to 625 resources from the French DBpedia [8] that describe historical monuments of Paris. The monuments in DBpedia are spatially referenced by prop-fr [9]:longitude and prop-fr:latitude properties. DBpedia is extracted from Wikipedia, a volunteered encyclopedia, where the location of the resources are provided without any metadata about their geometric modeling or their positional accuracy. Though, by plotting the data on a base map we can intuitively distinguish three principal representations of the monuments locations : near the building center, near the building facade or near the road centerline. The assumption that these are the actual intended geometric representation cannot be verified because of the open volunteered nature of this source. However, we take it as hypothesis in order to automatically learn the geometry capture rules and estimate the positional accuracy of the geometries.

We used two reference geographic data sources about buildings [10] and road network [11] and we investigated some possible learning features computed with state of the art GIS tools and presented in Fig.2.3. We prepared a training set by manually labeling ~30 monuments of each class. Then, we applied some of Weka [12] learning algorithms and validated the results manually (see results in table 2.1). The resulting classification is interpreted on the level of the geometries by adding metadata about their capture rules.

Finally, we estimated the absolute positional accuracy of each point by summing two values : the distance between the point and its intended location (the building facade, the building barycenter or the road centerline) and the planimetric accuracy of the geographic feature that represents this intended

8. `http://fr.dbpedia.org/`. Version of December 2013.
9. `http://fr.dbpedia.org/property/`
10. From the BD PARCELLAIRE®, IGN's land parcels database.
11. From the BD TOPO®, IGN's topographic database.
12. `http://www.cs.waikato.ac.nz/ml/weka/`

TABLE 2.1 : Learning results for each classifying algorithms

Method	Precision	Recall	F-measure
Bayes Network	91,6%	91,3%	91,3%
JRIP	96,3%	96,3%	96,3%
Decision Table	96,4%	96,3%	96,2%
Random Forest	98,8%	98,8%	98,7%

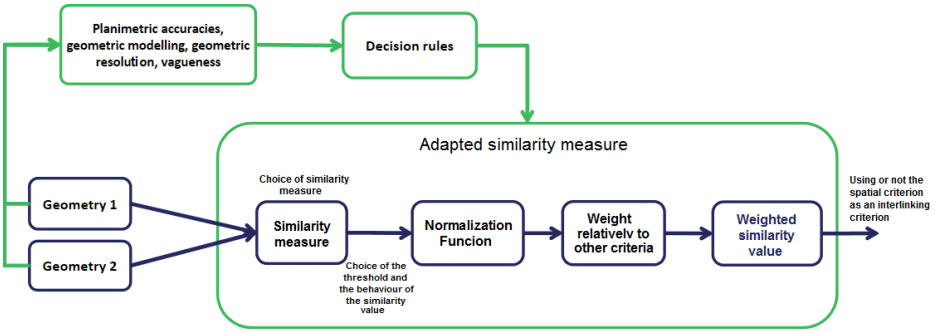

FIGURE 2.4 : The global approach for the self-adaptive comparisons of geometries

location. More details on the population approach described in this section are presented in [FAH17b]

2.3 An Adaptive Approach for Geometry Similarity Evaluation

In this section, we present how to use parameters about thresholds or the confidence value function, inferred using the XY semantics ontology, to automatically adapt the in-progress geometry similarity evaluation process.

2.3.1 General Description of the Approach

We have seen in section 2.1.1 that geometry similarity measures are usually based on some main choices : one or more similarity functions, the behavior of their normalization function, their parameters such as thresholds, the way they are combined through aggregation operators and weights. All these choices can be made based on the knowledge provided by the metadata about geometries capture process represented consistently with the XY semantics vocabulary.

As shown in Fig. 2.4, we suggest choosing the geometry distance function and customizing the confidence function for every comparison of geometries, depending on their metadata. For example, the threshold, the confidence function and the weight of the spatial criterion can be adapted. This can be decided through some decision rules that take as input the metadata of two geometries and give as output the parameters for their comparison. The decision rules must be defined by a data matching expert. The decisions concern different cases and can impact distinct parameters :

— The value of the distance threshold.

— The behavior of the confidence function.

— The weight of the spatial criterion.

— The potential neutrality of the spatial criterion.

For example, when two geometries have different capture rules, we can expect a considerable gap between them. In this case we can be less strict with distance values (i.e. define a higher distance threshold), and thus we can provide higher confidence values for the same distance. Moreover, when two geometries have a bad absolute positional accuracy, or when they are captured based on some vague geographic entity, we can increase or decrease the weight of the spatial criterion or even removing it from the aggregation, depending on its estimated reliability.

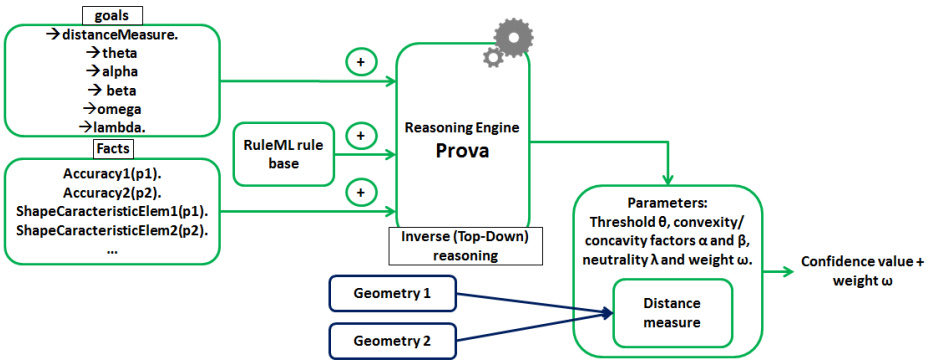

FIGURE 2.5 : Implementation of the self-adaptive linking approach

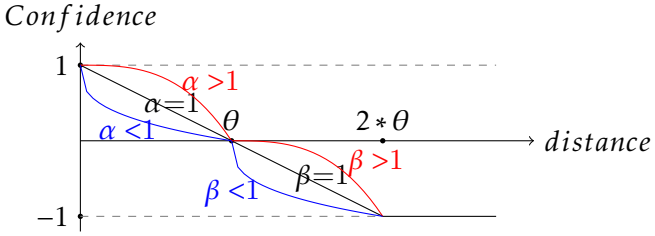

FIGURE 2.6 : Variations of confidence function behavior

2.3.2 Approach Implementation

Silk [13][IJB11] is a very well known and maintained data linking tool with many interesting features. In order to capitalize on the assets of Silk, we implemented our approach as described in Fig. 2.5 to make Silk compatible with our adaptive approach for geometry similarity evaluation.

In the case of Silk, the confidence value [14] is obtained as described in equation 2.1, where d is the distance computed between property values and θ is the threshold chosen by the user.

$$confidence = \begin{cases} 1-(d/\theta) & if \quad d \in [0, 2\times\theta] \\ -1 & else \end{cases} \tag{2.1}$$

As explained in 2.3.1, decision rules based on the metadata that describe the geometries may impact the confidence function behavior. We thus suggest changing Silk's default confidence function and replacing it by the following function :

$$confidence = \begin{cases} 1-(d/\theta)^{\alpha} & if \quad d \in [0, \theta] \\ -((d-\theta)/\theta)^{\beta} & if \quad d \in [\theta, 2\times\theta] \\ -1 & else \end{cases} \tag{2.2}$$

Where d is the distance computed between geometries and θ, α and β are the parameters affected by the decision rules. θ represents the threshold. It can be computed by summing the absolute positional accuracy values of the geometries. α and β are the convexity/concavity factors for respectively the positive and the negative parts of the confidence function. Both α and β should be positive values (c.f. Fig. 2.6). When the decision rules affect one or more of these three parameters, the confidence function becomes more or less strict while keeping its monotony. Two other parameters are eventually necessary : λ the neutrality of the spatial criterion and ω its weight in the case of a weighted aggregation.

All the parameters used by the geometry comparison process are defined by means of a decision rule base defined in advance by a data linking expert. The rule base is written in RuleML [15].The pieces of knowledge about geometries capture process and planimetric accuracy are formatted as RuleML facts. The resulting values of the requested parameters are represented as RuleML goals.

13. http://silkframework.org/
14. https://github.com/silk-framework/silk/blob/master/doc/LinkageRules.md
15. Rule Markup Language, http://wiki.ruleml.org/

Then, a backward reasoning is executed by PROVA [16] rule engine on the whole rule base to infer the comparison parameters. θ, α, β, ω and λ are thus given as an output from the reasoning engine. They are then used together with the spatial distance to compute the confidence value of the spatial criterion respecting the formula described by the equation 2.2. A backward (Top-Down) reasoning works similarly to most Prolog systems[BBH+05]. It gives answers to a set of goals by reasoning on a base of rules and facts. In order to make sure that our approach works safely, the rule base must be decidable. Since we have a fixed number of goals, the complexity of the reasoner is linear with the number of the rules declared in the rules base. Using PROVA in Silk for every comparison results in a total quadratic complexity. This implementation is scalable though, because it remains compatible with the MapReduce version of Silk.

2.4 Interlinking the Monuments of Paris

To evaluate our adaptive geometry similarity measure, we applied it on datasets on Paris historical monuments. The first one is the Mérimée [17]database, which is a national monuments registry produced and maintained by the French Ministry of Culture and Communication. Mérimée database contains 1582 instances, it is provided as a CSV file and its monuments are located by textual addresses. We transformed the data to RDF with the Datalift [18] platform. Then we selected only the monuments located in Paris and we geocoded these monuments by linking their addresses to their corresponding BD ADRESSE® [19] features. The second one comes from DBpedia and is presented in 2.2.3

Extracting metadata about the absolute positional accuracy and the geometry capture rules of Mérimée resources is quite straightforward. Indeed, their geometries come from the BD ADRESSE® database, which is a well-documented traditional geographic database. Its implementing rules provide information about the four different geometry capture rules applied for addresses : at the address sign, by projection on the corresponding street centerline, by interpolation on the corresponding section of the street centerline, and more rarely in the center of the addressing zone. Depending on these geometry capture rules, three different positional accuracy values are possible : 12, 18 and 30 meters. We structured these metadata according to the vocabulary presented in 2.2 and we added them to Mérimée geometries with Sparql update queries. In the case of DBpedia monuments, we used the approach presented in section 2.2.3 to learn their metadata.

2.4.1 Comparative Matching Tests

In order to evaluate our approach, we have performed two spatial matching tasks on the datasets described above : one with Silk's default spatial distance operator and one with our adaptive geometry similarity measure.

For Silk's default spatial distance operator, the confidence value of each comparison is computed according to the formula 2.1. We performed several runs with different distance thresholds θ to find out which value gives the best results. The top table of Table.2.2 outlines these results. The best f-measure is obtained for θ=40 meters. The runtime of this approach is around 2 seconds.

For our approach, we used two rule bases to define θ, α, and λ parameters (see section 2.3.1). Since it is a mono-criterion matching task, the parameters β and ω are not needed. Since the geometries in the two datasets are points only, we chose to use a euclidean distance measure between them. With respect to what is usually done in geographic data matching, we define the threshold θ as the sum of the positional accuracies of the two compared geometries (here named $accu1$ and $accu2$). When the geometries capture rules targeting different spatial characteristics ($characelem1$ and $characelem2$) of the real world entities they intend to represent (e.g. the first geometry is the barycenter of the building while the second is a point of the facade of a building), we add a bias named $delta_c$. When they target different types of real world entities ($host1$ and $host2$), we add another bias named $delta_h$. For our use case, we set $delta_c$ at 10 m and $delta_h$ at 15 m. The rule base $rb1$ is summarized below and the results we get with it are shown on

16. https://prova.ws/
17. https://www.data.gouv.fr/fr/datasets/immeubles-proteges-au-titre-des-monuments-historiques/
18. http://datalift.org/
19. Address database of IGN (French national mapping agency).

Table.2.2 :

```
theta(X):–host1=host2,characelem1=characelem2,X=accu1+accu2.
theta(X):–host1=host2,characelem1!=characelem2,X=accu1+accu2+delta_c
theta(X):–host1!=host2,X=accu1+accu2+delta_h.
distance(”euclidian”).
alpha(1). lambda(false).
```

Based on the experience of the test performed with the rule base $rb1$, we defined a second rule base by adding more fine-grained rules, namely $rb2$. In this rule base, we set the values of $delta_c$ at 20 m and $delta_h$ at 30 m [20]. We also change the convexity of the confidence function when geometry capture rules are different and when the targeted real world geographic entities are too vague or too wide, we neutralize the spatial criterion. These additional rules are described below and the results are also presented in Table.2.2. These rules replace the last line of $rb1$:

```
alpha(1):– host1=host2,characelem1=characelem2.
alpha(2):– host1=host2,characelem1!=characelem2.
alpha(3):– host1!=host2.
lambda(true):–host1= addressZone. lambda(true):–host2= addressZone.
lambda(true):–host1= commune. lambda(true):–host2= commune.
lambda(false):–host1!=addressZone,host2!=addressZone,host2!=commune,
              host2!=commune.
```

The runtime of the matching task using these rule bases is 9∼14s.

TABLE 2.2 : Instance matching results compared to a reference links set produced with Wikidata information and completed manually

θ	Precision	Recall	F-measure
10	84,55%	20,90%	33,51%
20	74,15%	39,33%	51,40%
30	64,15%	51,46%	57,11%
40	57,96%	58,88%	**58,42%**
50	50,09%	63,15%	55,86%
60	44,75%	65,17%	53,06%
70	40,43%	66,97%	50,42%
80	36,76%	68,31%	47,80%
90	33,99%	69,44%	45,64%
100	31,68%	70,34%	43,68%

Using Bayes Network learning results			
Rule base	Precision	Recall	F-measure
rb1	70,43%	58,88%	64,14%
rb2	71,43%	62,92%	**66,91%**
Using learning results after correction			
Rule base	Precision	Recall	F-measure
rb1	73,46%	59,10%	65,50%
rb2	70,99%	62,70%	**66,59%**

20. $delta_c$ and $delta_h$ were estimated by investigating the bias in some cases where two geometries with different geometric modelings locate two equivalent resources

2.4.2 Discussion

Extracting knowledge about the causes of heterogeneity between geometries by using supervised learning method shows promising results. Nonetheless, the choice of adequate learning features is conditioned by the context of the data and has an important impact on the results. For instance, we ran another learning test on the address data in the city of Lyon in France. In this case the geometries had also three possible geometry capture rules : in the center of the building, at the entrance of the building and on the street centerline. In this city, the buildings sizes and their distances to the road are more homogeneous than in Paris. In this case, simple learning features such as d_f, d_r and d_c (Fig.2.3) were sufficient to obtain very good results. The choice of the training set is also crucial : the entities must be clearly representative of the different learning classes. The classification errors induced by the learning step show a low effect on the final linking results compared to the improvement brought by the linking approach.

The matching results of our approach show clearly better f-measure scores than the default approach. Adapting the parameters of the geometry comparison measure ensures some of the benefits of both small and big distance thresholds. Compared to the best result of the classical approach (θ=40), we avoid 50% of the false positive links using $rb1$ and 40% using $rb2$. We do not significantly add new true positive links using $rb1$ but we increase their number by 6% using $rb2$.

Unsurprisingly, our approach has a clearly higher runtime. The complexity of the implementation depends on the size of the rule base. This is why we have tried to define the minimum number of decidable rules that can sufficiently adapt the parameters of the confidence function. A more detailed rule base could have provided better results but it would have been much less efficient in runtime. The user has to find the best trade-off between efficiency and performances. As a matter of fact, our approach is better suited to instance matching tasks of data sources which have a high spatial density and instances described by geometries with a lot of internal geometric heterogeneities.

2.5 Conclusion

In this chapter we tackled the problem of the geometry similarity evaluation for linking georeferenced resources. We proposed an ontology to represent knowledge about geometry positional accuracy and capture rules. We also proposed an approach to extract it from the considered spatial data and geographic reference data by using automatic supervised learning. We also defined a data matching approach that relies on this knowledge to adapt the comparison of geometries during its runtime. The matching results show better performances than the classical non-adaptive approach.

Yet, the main downside of our approach is the time complexity of the current implementation that should be improved. This could be done by adding a cache system for the reasoning results in order to reduce the workload of the reasoning engine. Further tests, with bigger and more heterogeneous datasets, especially datasets with different types of geometry, could also bring new insights to this proposal. Future tests should also include the two remaining aspects of the XY semantics, namely the geometry resolution and its vagueness, in both populating and interlinking approaches.

Chapitre 3

Enriching KGs with Contextual Identity Links

This chapter is based on work realized in the context of the PhD thesis of Pierre-Henri Paris that I co-supervised with Dr. Samira Cherfi (from Cnam Paris). The result of this work was published in : [PHC20c], [PHC19a], [PHC20a], [PHC20b], [PHC19b]

Open and RDF-based knowledge graphs (KGs), like prominent Wikidata [1] or DBpedia [2], are continuously growing in terms of size and usage. Consequently, the number of entities described in those KGs leads to a problem for both data publishers and data users : **how to know if two entities are the same or not ?** According to Noy et al. [NGJ+19], this question remains one of the top challenges in knowledge graphs industry. To interlink KGs, the *owl:sameAs* property has been defined by the W3C [3] in 2004 to link entities that are allegedly the same. Indeed, a (real world) object is described among several KGs, and those descriptions are linked thanks to the *owl:sameAs* property. However, the semantic definition of *owl:sameAs* is very strict. It is based on Leibniz's identity definition, i.e., the identity of indiscernibles.

Hence, two entities are considered identical if they share all their $\langle$ *property,value* $\rangle$ couples in all possible and imaginable contexts. In other words, two entities are identical if **all their properties are indiscernible** for each value. Once an identity link is stated between two entities, it is possible to use $\langle property, value \rangle$ couples from one entity to another. However, it is a very strong assertion to state that two objects are the same whatever the context. From a philosophical point of view, there are multiple counter-arguments to the definition of Leibniz's identity. For example, if we consider two glasses from the same set of glasses, they are indiscernible from each other and yet they are two different physical objects. Is a person the same as she was ten years ago ?

It is also a technical problem because of the open-world assumption [DS06], on the one hand, and on the other hand, because of what a data publisher has in mind that could be different from what the user expects when using data. Besides, when data is published, it is "almost" impossible to know the **consensus** behind the decision of creating an *owl:sameAs* link. Several works ([HHM+10] or [DSFM10]) have demonstrated that the use of *owl:sameAs* was inadequate. Indeed, established links might be considered as true only in specific contexts. According to [NGJ+19], the problem of identity management in knowledge graphs remains one of the top challenges in the industry.

As a first intuition, a contextual identity between two entities might be seen as a subset of properties Π for which these entities share the same values for each $p \in \Pi$.

Example 3.0.1 : Two different generic drugs *Drug1* and *Drug2* can be identical when considering the active ingredient. If a Knowledge Graph contains the triples $\langle$ *Drug1 activeIngredient Molecule1* $\rangle$ and $\langle Drug2\ activeIngredient\ Molecule1 \rangle$, then $Drug1 \equiv_{activeIngredient} Drug2$ when the context is *activeIngredient*.

One of the core features of *owl:sameAs* is to be able to **propagate all properties** from an entity to other identical entities. Hence, *owl:sameAs* allows discovering more knowledge and to increase completeness. In the same way, contextual identity must help to discover **more knowledge and to increase completeness**, but only under specific circumstances. So, to be useful, a contextual identity must specify what is happening with properties that are not part of the context. In other words, **an identity context must have propagable properties**.

Example 3.0.2 : Following the example 3.0.1, stating only $Drug1 \equiv_{activeIngredient} Drug2$ has a limited interest, if we do not know what to do with other properties besides $activeIngredient$. Considering the context $activeIngredient$, the property $targetDisease$ is propagable, and if the statement $\langle$ *Drug1 targetDisease Disease1* $\rangle$ exists then we can state that

1. https://www.wikidata.org
2. https://wiki.dbpedia.org/
3. https://www.w3.org/TR/owl-ref/

⟨*Drug*2 *targetDisease Disease*1⟩. But if we consider the property *excipient*, then it is not propagable.

Moreover, the ability to propagate a property between entities depends on the context, i.e., the same property might be propagable in a context C_1 and not propagable in a context C_2.

Several works have attempted to propose a solution to the contextual identity. [BSvH16], [IHvH+17] and [RPS17] defined three different ways to handle identity under a given context. However, none of those works propose a solution to discover properties that can be propagated given a specific context.

Research questions : With a given identity context between two entities, how to find properties that can be propagated ? Is it possible to find propagable properties (semi-)automatically ?

In this chapter, based on the context definition of [IHvH+17], we propose an approach to **find propagable properties** to facilitate knowledge discovery for users. Instead of manually listing the propagating properties as in [IHvH+17], we automatically identify the propagating properties for a given context using semantic textual similarity, significantly reducing burden to users. The semantic similarity is based on the sentence embeddings corresponding to the textual descriptions of the properties. Our intuition is inspired by Tobler's first law [Tob70], that is :

"*Everything is related to everything else, but near things are more related than distant things.*"

Therefore, **we hypothesize that, from a semantic point of view, the closer a property is to the identity context, the more likely it could be a right candidate for propagation**. So, the idea is to **compute a distance between indiscernible properties** and **candidate properties for propagation**. Consequently, numbers, and in our case numerical vectors, are best suited to compute this distance. A numerical representation of the textual description of each property through its *rdfs:comment*[4] or *schema:description*[5] can provide a basis to get this vector. Indeed, sentence embeddings of properties descriptions give us numerical vectors which distributions in the vector space comply with the semantic similarity of the sentences. We validated our approach through quantitative and qualitative experiments.

The rest of the chapter is organized as follows. In the following sections we present the related work. Then, in Section 3.3, we present our approach. In Section 3.4, we present the quantitative and qualitative experiments we have conducted.

3.1 State of the Art

In the first part of this section, we describe papers that pointed out the problems raised by the *owl:sameAs* usage. In the second part, we discuss the proposals that tackle these problems.

3.1.1 Identity Crisis

As early as 2002, Guarino and Welty [GW02] raised the issue of identity for ontologies. Especially when time is involved, stating that two things are identical became a philosophical problem. The authors proposed to involve in identity only essential properties, i.e., a property that cannot change. As described in Horrocks et al. [HKS06], the *owl:sameAs* property purpose is to link two entities that are strictly the same, i.e., both entities are identical in every possible context. *owl:sameAs* has a strict semantics allowing to infer new information. Many existing tools produce such *owl:sameAs* links [FNS11], and several surveys are available to this end [FNS11, ABT16, NHNR17].

However, none of these approaches consider contextual identity links. Their purpose is to discover identity links that allegedly always hold. This is, from a philosophical point of view, hard to obtain as underlined by Leibnitz's identity definition. Indeed, as stated for example in Halpin et al. [HHM+10] or Ding et al. [DSFM10], because of the strict semantic of *owl:sameAs*, the burden of data publishers might be too heavy. As a matter of fact, *owl:sameAs* links are not often adequately used. Some might be simply wrong, and, more insidiously, some might be context-dependent, i.e., the *owl:sameAs* link does not hold in every possible context because it is hard to obtain a consensus on the validity of a statement. What a data modeler means may not be what a data user expects. This misuse of *owl:sameAs* is often referred to

4. https://www.w3.org/TR/rdf-schema/#ch_comment
5. https://schema.org/description

as the "identity crisis" ([HHM+10]).

3.1.2 Contextual Identity

Very few works have addressed the contextual identity problem. Beek et al. [BSvH16] addressed this issue by constructing a lattice of identity contexts where contexts are defined as sets of properties. All entities belonging to a context share the same values for each property of this context. Hence, a context is a set of indiscernible properties for an entity. However, the authors do not give indications about the usage of properties not belonging to such contexts. Raad et al. [RPS17] proposed an algorithm named DECIDE to compute contexts, where identity contexts are defined as sub-ontologies. Nevertheless, as in the first work, properties of entities that are not in the sub-ontology are ignored. So, in both previous works, there is a limitation of properties that do not belong to a context. This limitation cripples the interest of using such approaches. Indeed, one of the goals of an identity context is to define an identity relation between two entities to use information about one on the other. The solution by Idrissou et al. [IHvH+17] involves such propagation of properties, and thus, increases completeness of an entity according to a context. However, this proposal requires users to provide both the propagating and indiscernible properties as input. Hence, it leaves the burden to the user to identify and provide context and properties. The user must provide the two sets of indiscernible and propagating properties.

In this work, we propose to remove this burden partially from the user, i.e., to **semi-automatically compute the propagation set of properties given an indiscernibility set of properties**. For this, we will use sentence embedding (presented in Section 3.3.3) to compute the embeddings of properties using their descriptions to discover the **propagating properties** with respect to a given identity context (as defined in [IHvH+17]).

3.2 Motivation

Sometimes, real-world entities may be close regarding their properties but not exactly the same. For example, the French capital, Paris, is both a city and a department (an administrative subdivision of the French territory). While considering that the city and the department are the same concerning their geography, they are two distinct entities administratively (or legally) speaking. Now, suppose both Paris are represented in a Knowledge Graph as distinct entities, and both are linked to (possibly distinct) movie theaters. If one wants to retrieve movie theaters located in the city of Paris, results will not be complete if some of them are linked to the department (see Figure 3.1).

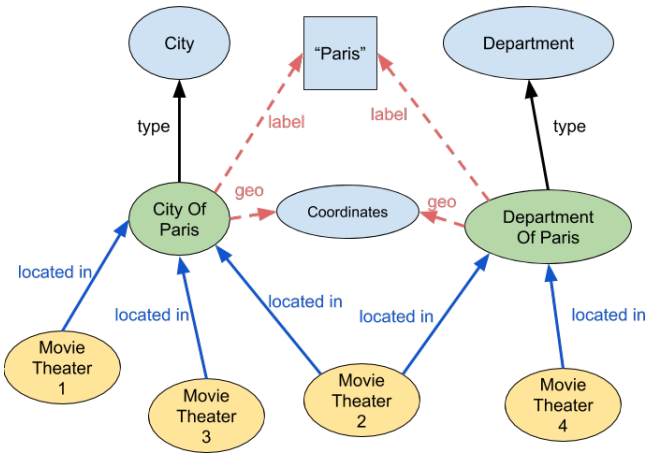

FIGURE 3.1 : Excerpt of a Knowledge Graph about Paris, France. The properties in red are indiscernible for both the city and the department. The properties in blue are propagating given the red properties are indiscernible.

A French citizen might know this ground truth, but how to allow an automated agent to discover this

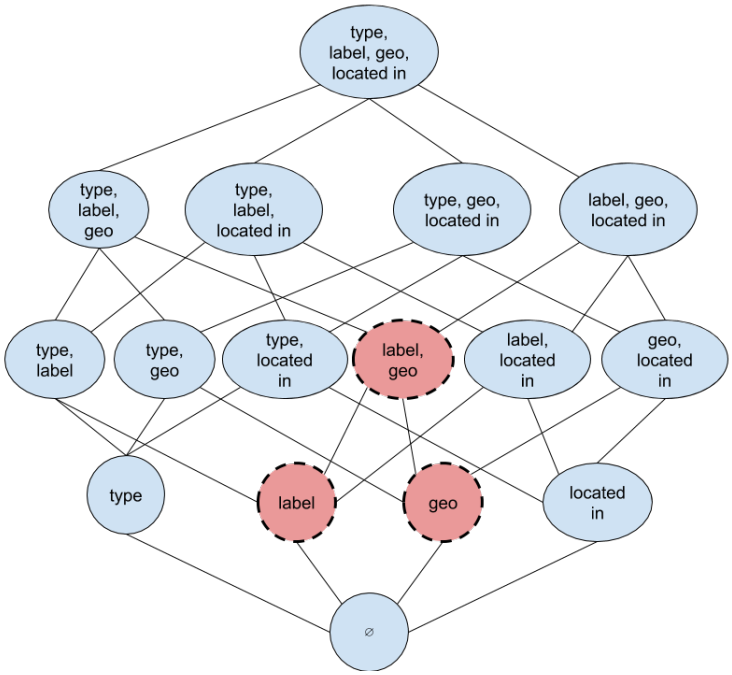

FIGURE 3.2 : Simplified identity lattice from Figure 3.1 : each node is an indiscernible set of properties. Only the red nodes have similar entities.

fact ? Contextual identity is a possible answer to this question, i.e., a set of properties for which values are the same for both entities. Considering the present example, both Paris (city and department) are geographically the same, and some properties related to geography might be **propagated**. In Figure 3.1, the red properties (*geo* and *label*) are indiscernible (have the same values), and the blue properties (*located in*) are propagating. In the real world, movie theaters located either in the city or the department, according to the Knowledge Graph, are located in the same place. Although the two entities do not share the same values for the *located in* property, this one is related to the geographic context. Indeed, for a human agent, the *located in* property might be obviously propagated between the two entities.

While we expected to have the four movie theaters located in Paris, the query in Listing 3.1 will only return movie theaters 1, 2, and 3 (see Figure 3.1).

```
SELECT DISTINCT ?movieTheater WHERE {
    ?movieTheater :locatedIn :CityOfParis .
}
```

Listing 3.1 : SPARQL query retrieving all movie theaters in Paris, France.

Thus, discovering such contexts of identity between entities might improve the completion of query results. Our intuition is inspired by Tobler's first law ([Tob70]), that is :

"*Everything is related to everything else, but near things are more related than distant things.*"

Therefore, **we hypothesize that, from a semantic point of view, the closer a property is to the identity context, the more likely it could be a right candidate for propagation**. In the previous example, *located in* clearly refers to a geographic fact, and the context of identity is about geography since it is composed of geographical coordinates. So, the idea is to **compute a distance between indiscernible properties** and **candidate properties for propagation**. Consequently, numbers, and in our case, numerical vectors, are best suited to compute this distance. A numerical representation of the textual description of each property through its *rdfs:comment* or *schema:description* can provide a basis to get this vector. Indeed, the embedding of property descriptions gives us numerical vectors whose distributions in vector space respect the

semantic similarity of sentences.

3.3 Approach

In this section, before diving deeper into the core approach, we give some definitions needed later to describe the approach.

3.3.1 Preliminaries

We first need to formalize the definition of a propagable property.

Definition 3.3.1 : (**Propagable Property**) The property p can be propagated from an entity e_1 to an entity $e_2 \leftrightarrow (\forall o : \langle e_1, p, o\rangle \rightarrow \langle e_2, p, o\rangle)$.

Several propositions have been made to define an identity context. We choose the one from [IHvH+17] since it is the only one that considers the propagation of properties. They give the following definition of the identity context :

Definition 3.3.2 : (**Identity Context**) An identity context $\mathcal{C} = (\Pi, \Psi, \approx)$ is defined by two sets of properties (Π and Ψ) and an **alignment procedure** ($\approx$). Π is the **indiscernibility set** of properties (equation 3.1) and Ψ is the **propagation set** of properties (equation 3.2). In the following, x and y are entities.

$$\begin{aligned} x =_{(\Pi,\Psi,\approx)} y \leftrightarrow \forall (p_1, p_2) \in \Pi^2 \text{ with } p_1 \approx p_2 \\ \text{and } \forall v_1, v_2 \text{ with } v_1 \approx v_2 : \langle x, p_1, v_1\rangle \leftrightarrow \langle y, p_2, v_2\rangle \end{aligned} \tag{3.1}$$

$$\begin{aligned} x =_{(\Pi,\Psi,\approx)} y \rightarrow \forall (p_1, p_2) \in \Psi^2 \text{ with } p_1 \approx p_2 \\ \text{and } \forall v_1, v_2 \text{ with } v_1 \approx v_2 : \langle x, p_1, v_1\rangle \leftrightarrow \langle y, p_2, v_2\rangle \end{aligned} \tag{3.2}$$

Moreover, we define the **level of a context** $|\Pi_{\mathcal{C}}|$ as the number of its indiscernible properties.

In the case where similar entities according to an identity context belong to the same Knowledge Graph, it is not necessary to have an alignment procedure.

An entity can have several identity contexts, depending on properties in the indiscernibility set Π. Indeed, two different combinations of properties can give different sets of similar entities. The identity lattice of all identity contexts of an entity e is defined as follows (see Figure 3.2) :

Definition 3.3.3 : (**Identity Lattice**) An identity lattice $\mathcal{L}$ is a lattice, where each element is an identity context. The set inclusion between indiscernibility set of properties of each context is the binary relation responsible for the partial order.

The last notion is the seed of a lattice or a context that we define as follows :

Definition 3.3.4 : (**Seed of a lattice or a context**) Each context of a lattice is constructed from the **same entity** e for all the contexts of the lattice. This entity e is called the seed of the lattice.

Indeed, to build an identity lattice, we need to start from a seed, despite the fact that the lattice could potentially be valid with another seed (see Figure 3.2).

Now that we have defined the necessary concepts, we will explain the core of our approach.

3.3.2 Computation of contexts

We present Algorithm 1 that computes an identity lattice. It takes as input the seed entity, the source KG to which the seed belongs, the target KG (possibly the same as the source KG) and an alignment procedure if the two KGs are distinct. The main idea is to start by computing level one identity contexts with each seed's property and finally combine those contexts to obtain upper-level identity contexts. When building a context, its first part is its indiscernibility set, from which we then get similar entities, to obtain candidate properties for propagation and, in the end, propagating properties.

The first step, line 5, is to compute all level 1 identity contexts (see Definition 3.3.2). Indeed, for each property p of the seed, there is exactly one identity context (its indiscernibility set is $\Pi = \{p\}$). Later,

Data : $\mathcal{KG}_1$: the source KG, $\mathcal{KG}_2$: the target KG, *seed* : an entity of $\mathcal{KG}_1$, $\approx$: an alignment procedure between $\mathcal{KG}_1$ and $\mathcal{KG}_2$
Result : $\mathcal{L}$: a lattice of identity contexts between the seed and entities in the target KG

1 $\mathcal{L} = \emptyset$;
2 /* Get all explicit and implicit types of the seed */
3 $\mathcal{T}_{seed} = \{t : \langle seed\ rdf{:}type\ t\rangle \in \mathcal{KG}_1\}$;
4 /* the following will create all contexts of the level 1 (with only one indiscernible property) */
5 **for** *each property p of seed* **do**
6 $candidateEntities = \emptyset$;
7 **for** *each value o such as* $\langle seed\ p\ o\rangle \in \mathcal{KG}_1$ **do**
8 /* $entities_{p,o}$ is the set of indiscernible entities with *seed* with respect to the p,o pair */
9 $entities_{p,o} = \{e : (\exists (p', o'), p' \approx p, o' \approx o, \langle e\ p'\ o'\rangle \in \mathcal{KG}_2) \wedge (\exists t \in \mathcal{T}_{seed}, t' \approx t, \langle e\ rdf{:}type\ t'\rangle \in \mathcal{KG}_2)\}$;
10 **if** $entities_{p,o} \neq \emptyset$ **then**
11 $candidateEntities = candidateEntities \cup \{entities_{p,o}\}$;
12 **end**
13 **end**
14 /* $entities_p$ is the set of indiscernible entities with *seed* with respect to the property p */
15 $entities_p = \bigcap candidateEntities$;
16 $\Psi = getPropagationSet(seed, entities_p, \{p\})$;
17 **if** $\Psi \neq \emptyset$ **then**
18 $\Pi = \{p\}$;
19 $\mathcal{C} = (\Pi, \Psi, \approx)$;
20 $\mathcal{L} = \mathcal{L} \cup \mathcal{C}$;
21 **end**
22 **end**
23 /* Now we can combine contexts of the same level */
24 **return** $constructUpperLevels(\mathcal{L}, \mathcal{KG}_1, \mathcal{KG}_2, seed, \approx)$

Algorithm 1 : createLattice : calculate identity lattice of an entity.

identity contexts with only one indiscernibility property will be merged to give identity contexts of higher-levels. Next, we retrieve similar entities $entities_p$ to the seed that have the same value(s) for the given property p. If p is multi-valuated, then entities in $entities_p$ are similar to the seed for all values o such that $\langle seed\ p\ o\rangle$. It is worth noting that, when filling $entities_p$, we search only entities that have the same type(s) with the seed. This is because we want to avoid absurd results, e.g., comparing a person with an airplane. It also has the advantage of lowering the number of possible identity contexts to compute. Finally, based on $entities_p$, we compute the propagation set Ψ (line 10) as explained in the following section (Section 3.3.3).

The second step (see Algorithm 2) is to compute upper-level identity contexts based on those from level 1. The loop (line 3) of the algorithm calculates these upper-levels by combining contexts of the same level, and stops when it cannot construct new upper-level identity contexts. This calculation is based on an identity lattice operator, which is the set inclusion on indiscernibility sets. For example, a level 2 context is built on two contexts from level 1. Again, to lower the number of possible identity contexts to compute, if there is no similar entity to the seed for a given context C_i, there is no need to compute higher-level contexts based on C_i.

3.3.3 Propagation set using sentence embedding

Our approach for computing propagation set (Line 10 in Algo. 2) is elaborated in Algo. 3. It is based on sentence embedding which maps a sentence to a numerical vector. Ideally, semantically close sentences appear nearby in the numerical vector space.

Sentence embedding is a technique that maps a sentence to a numerical vector. Ideally, semantically close sentences are represented by close vectors in the numerical space considered.

Data : $\mathcal{L}$: the lattice with only level one contexts, $\mathcal{KG}_1$: the source KG, $\mathcal{KG}_2$: the target KG, *seed* : an entity of $\mathcal{KG}_1$, $\approx$: an alignment procedure between $\mathcal{KG}_1$ and $\mathcal{KG}_2$
Result : $\mathcal{L}$: a lattice of identity contexts between the seed and entities in the target KG

```
1  /* lvl is the current level in the lattice                                   */
2  lvl = 1 ;
3  while ∅ ∉ L do
4      contexts = ∅ ;
5      for (C1, C2) ∈ {(Ci, Cj) ∈ L × L : |ΠCi| = |ΠCj| = lvl, i > j} do
6          Π = ΠC1 ∪ ΠC2 ;
7          /* getEntities function gives the set of entities that are similar under the given
                identity context in the given KG                                */
8          entities = getEntities(C1, KG2) ∩ getEntities(C2, KG2) ;
9          if entities ≠ ∅ and Π ∉ L then
10             Ψ = getPropagationSet(seed, entities, Π) ;
11             /* see Algo. 3                                                   */
12             if Ψ ≠ ∅ then
13                 C = (Π, Ψ, ≈) ;
14                 contexts = contexts ∪ C ;
15             end
16         end
17     end
18     L = L ∪ contexts ;
19     lvl = lvl + 1 ;
20 end
21 return L
```

Algorithm 2 : constructUpperLevels : calculate upper-levels of the identity lattice of an entity.

Example 3.3.1 : "A soccer game with multiple males playing" and "Some men are playing a sport" are semantically close, thus their vectors should be close in terms of distance.

Reciprocally, two sentences that are not semantically related should have distant vectors.

Example 3.3.2 : "A man inspects the uniform of a figure in some East Asian country" and "The man is sleeping" should have distant vectors.

Those vectors enable usage of various mathematical operators that are obviously not available with chains of characters. One of the first major work in that field is *Word2Vec* [MCCD13] which captures co-occurrence of words. Each word is processed atomically and provides an embedding through two possible and distinct approaches, namely Skip–Gram and Continuous Bag of Words (CBOW). While CBOW aims is to predict a word given its context (i.e., previous and following words in a sentence), Skip–Gram will try to predict words with which a word is usually seen. Similarly, *GloVe* [PSM14b] provides embeddings for single words and might use Skip–Gram or CBOW. But *GloVe*, instead of capturing co-occurrence, focuses (in the end) on the count of appearance among contexts (i.e., previous and following words in a sentence). Then, *fastText* [BGJM17] is an extension of *Word2Vec* that treats words as n-gram of characters instead of as an atomic entity. N-grams sizes depend on input parameters. N-grams usage allows a better understanding of small words. Each n-gram is mapped to a vector and the sum of these vectors is the representation of the word. Another advantage of *fastText* is its capacity to provide an embedding even for unknown words, thanks to n-grams usage. While the three previous works are best suited to work with atomic words, the following computes embedding for a whole sentence.

The reasons behind using sentence embedding instead of a more classical distance measures, e.g., the edit distance, RDF graph embeddings like RDF2Vec [RP16], or an ontological alignment technique are : *(i)* classical string distances ignore sentence semantics, *(ii)* RDF graph embedding techniques are not yet adapted to such a task, and *(iii)* ontological alignment techniques align pairwise properties and not sets of properties.

Sentence embedding is widely used in several tasks such as computing semantic similarities between

Data : *seed* : the entity that generated Π,
entities : set of entities similar to *seed* with respect to Π,
Π : an indiscernibility set
Result : Ψ : a propagation set

```
/* computation of the embeddings of each property in Π by using one of the
   encoder                                                                  */
indiscernibilityEmbeddings ← getEmbeddings(Π);
meanVector ← mean(indiscernibilityEmbeddings);
/* getCandidateProperties function returns the set of all candidate properties for
   propagation                                                              */
candidates ← getCandidateProperties(Π, {seed} ∪ entities);
/* then compute their embeddings                                            */
candidatesEmbeddings ← getEmbeddings(candidates);
Ψ ← ∅;
for candidateVector in candidatesEmbeddings do
    similarity ← cosineSimilarity(candidateVector, meanVector);
    if similarity ≥ threshold then
        Ψ ← Ψ ∪ {candidateVector};
    end
end
return Ψ
```

Algorithm 3 : getPropagationSet : calculate the propagation set.

two texts. An encoder derives sentence embeddings, to capture the semantics of a language, from a large text corpus. A lot of attention has been given to sentence embeddings lately. Approaches like *Universal Sentence Encoder* [CYK+18], *GenSen* [STBP18] and *InferSent* [CKS+17] are among the state-of-the-art encoder for sentence embeddings. *InferSent*, proposed by [CKS+17], is a state-of-the-art encoder proved to be effective on sentence embedding. To train their supervised sentence embeddings model, the authors used the Stanford Natural Language Inference (SNLI) dataset that consists of more than 500K pairs of English sentences manually labeled with one of three categories (entailment, contradiction and neutral). They tested several architectures and find out that a BiLSTM network with max pooling offered the best results. A BiLSTM network is a bi-directional LSTM often used for sequence data, i.e., a recurrent neural network (with loops). Max pooling is a technique that allows reducing the number of parameters of the model by selecting the maximum value of a moving "window". Moreover, the pre-trained model is based on fastText, thus allowing computing meaningful representations even for out-of-vocabulary words, i.e., words that did not appear in the training data. *GenSen* [STBP18] and *Universal Sentence Encoder* [CYK+18] are both based on multi-task learning (MTL). MTL purpose is to learn multiple aspects of a sentence by switching from different tasks like translation or natural language inference. The former uses a bi-directional Gated Recurrent Units (GRU), which is a recurrent neural network like LSTM but with fewer parameters. The latter uses the transformer architecture that transforms a sequence into another but without recurrent neural network (unlike *InferSent* and *GenSen*). In Section 3.4.2, we will present results with those three encoders.

As presented in the beginning of this chapter, our intuition, based on Tobler's first law, is that the propagation set of properties can be found given an indiscernibility set, if vectors of descriptions of those two sets are close enough. As presented in Section 3.3.3, sentence embedding allows us to represent a sentence, i.e., a sequence of words, as a numerical vector. When two sentences are semantically close, their respective vectors should also be close in the considered space. In this work, we propose to use property descriptions (e.g., *rdfs:comment* or *schema:description*) as "*standard plug type for mains electricity in a country*") to find properties that are semantically related and consequently right candidates for propagation for a given indiscernibility set Π. For example, in Wikidata, the property "director" has the follow description :

"director(s) of film, TV-series, stageplay, video game or similar". Descriptions are mainly composed of one sentence. Most of the properties are described with such annotations, e.g., properties of Wikidata are annotated with an english *schema:description* at 98.9%. For the embedding computation, any of the previously described encoders can be used. We will provide in Section 3.4.2, an analysis of the different results obtained by these encoders.

The last algorithm presents our proposition to compute Ψ given a Π. It takes as input three parameters : a seed (an entity), a set of property built from the seed (indiscernibility set Π), and a set of entities that are similar to the seed with respect to Π. The computation of Π is presented in the previous section (see algorithm 1).

First, for each property in the indiscernibility set Π, we calculate its representational vector (see line 2). Then, we compute the weighted mean vector that represents the indiscernibility set (line 3). With use as weights the weight of properties as defined in that we proposed in [PHC19a]. Indeed, as explained, some properties are more important to determine identity. In an identity context, the indiscernibility set Π is a set of properties, thus we can compute for each property in Π its vector (see line 2). Then we can compute the mean of the vectors and have a numerical representation of Π (it is also a vector of the same size). This vector representing the mean of vectors derived from Π properties is noted $\sqsubseteq_{\Pi}$ in the following. Similarly, we consider each property of the seed or its similar entities and compute their representational vectors. Therefore, on the one hand, we have one vector that represents the set of indiscernibility and, on the other hand, we have n vectors for the n properties that are candidates for propagation. Properties of similar entities (with respect to the indiscernibility set Π) are also considered as candidates since possibly one of them can have a propagating property that the seed does not have (see line 5).

Then we loop on each candidate property to compute a cosine similarity [Sin01] between each candidate vector and the mean vector representing the indiscernibility set Π (line 9). If the cosine similarity is high enough (above a specified threshold as explained in the following section) the candidate property is considered as a propagable property.

Now that our approach has been presented, we will introduce experiments to validate our work.

3.4 Experimental Results

To evaluate our approach, we first implemented our approach and then conducted two types of experiments. In the first experiment, we built a gold standard upon Wikidata. Then, we computed standard precision, recall and F-measure against this gold standard. In the second experiment, we present several SPARQL queries that benefited from our approach.

3.4.1 Implementation and set-up

We implemented our approach in Python. For the sake of reproducibility, the code is made available on a GitHub repository [6]. As mentioned earlier, we used three sentence embedding approaches, namely *InferSent* [7], *GenSen* [8] and *Universal Sentence Encoder* [9]. We used an HDT file (see [MGF12] and [FMG+13]) that contains a dump of the last version of Wikidata [10]. HDT is a compressed serialization format for RDF graphs that allows a better reproducibility than a live SPARQL endpoint. Unlike Turtle or N-Triples, thanks to compression, HDT facilitates the manipulations needed to reproduce the experiments. The computer we used has an i7 processor and 32 GB of RAM. As an indication, the complete calculation of the identity lattice for an entity such as the city of Paris, France takes about 1396 ms. It has more than 1000 property-object couples and, in Wikidata, the mean number of property-object couples is about 60. Thus, it is a rather large entity and this approach could scale well.

6. https://github.com/PHParis/ConProKnow
7. https://github.com/facebookresearch/InferSent
8. https://github.com/Maluuba/gensen
9. https://tfhub.dev/google/universal-sentence-encoder/2
10. http://gaia.infor.uva.es/hdt/wikidata/wikidata2018_09_11.hdt.gz

3.4.2 Quantitative Study

The purpose of the quantitative experiment is to allow comparison with future approaches that may arise. To the best of our knowledge, our approach is the only one that has focused on the propagation of properties for contextual identity. Thus, one of the most important contributions of our work is the gold standard we provide. Indeed, it is not obvious, even for a human agent, to determine properties that might be propagated for a given indiscernibility set of properties.

Gold standard

As mentioned previously, we want to evaluate if our approach identifies **relevant propagable properties** to the user **according to a given context**. For this, we built a gold standard from the Wikidata Knowledge Graph that is known for its high data quality. It is also one of the most important actors of Linked Open Data initiative and is linked to many other knowledge graphs, e.g., DBpedia. Obviously, in this case, we no longer need an alignment procedure ($\approx$), since we do not consider multiple knowledge graphs (source and target knowledge graphs are the same).

We built 100 identity contexts, each context containing the indiscernibility set of properties Π and the propagation set of properties Ψ. We choose 5 classes (20 entities by class) : country, comics character, political party, literary work and film. Those classes have been chosen for several reasons. Firstly, they are sufficiently different to challenge our approach. Secondly, because the experts must judge which properties are propagable, it is easier for them if they have a minimum of knowledge about the subjects. Finally, it allows us to further investigate if results are different for different classes. For each selected class, we randomly selected one entity. Then we computed its identity lattice.

As stated before, the most difficult part when building the gold standard is to obtain a consensus among experts. A set of properties representing the indiscernibility set Π, and the experts must accordingly choose propagable properties in a set of candidates. It is the most time-consuming part to build the gold standard.

Execution against the gold standard

For each entity in the gold standard, we retrieved its partial identity context from the gold standard, i.e., the context with only the indiscernibility set Π. Then, we applied our algorithm on each set of indiscernibility to find the corresponding propagation set Ψ. For each context, we calculate the precision, recall, and F-measure as follows : True positive (tp) is the number of selected properties (by our approach) that are actually in Ψ, false positive (fp) is the number of selected properties (by our approach) and actually not in Ψ, false negative (fn) is the number of properties in Ψ not selected by our approach, $Precision = \frac{tp}{tp+fp}$, $Recall = \frac{tp}{tp+fn}$, and $Fmeasure = \frac{2 \times Precision \times Recall}{Precision+Recall}$ We then aggregated precisions, recalls and F-measures of each context thanks to the standard mean.

As there are no other approaches, to the best of our knowledge, retrieving candidate properties for propagation with respect to an indiscernibility set Π, we compared our approach with a baseline method. Instead of computing embeddings of descriptions, we computed the Jaccard index (JI) [Jac08] of property descriptions. Since the JI is a metric distance, we can compute the standard mean of several JI values and still get a distance. Moreover, the corresponding similarity is defined as $distance = 1 - similarity$ and both distance and similarity are normalized, thus the similarity can be used in the same way as with cosine similarity of embedded vectors.

The first series of results is presented in the Figure 3.3. It shows the evolution of the average precision of the baseline (in blue), *InferSent* (in red), *GenSen* (in yellow) and *Universal Sentence Encoder* (in green). The threshold has been tested from 0 to 1 with steps of 0.05. For each threshold, we computed the standard mean of the precision for the 100 entities of the five classes (country, comics character, political party, literary work and film). The goal is to evaluate the suitability of the different embedding technic and the baseline. At first glance, we observe that the four methods produce the same results for very low thresholds with precisions in average at 0.3. This is because the median number of candidate properties is 12 and the median number of propagable properties (in *Psi*) is 3. Therefore, even if all properties were

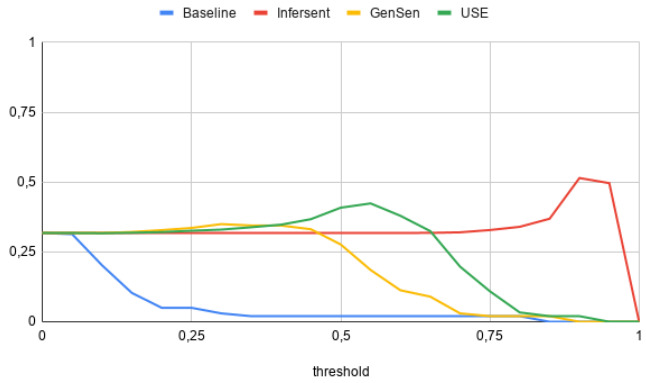

FIGURE 3.3 : Comparison of the average precision by thresholds for all five classes (country, comics character, political party, literary work, film). The threshold takes values from 0.5 to 0.95 by steps of 0.05. The baseline is in blue, InferSent in red, GenSen in yellow and Universal Sentence Encoder in green.

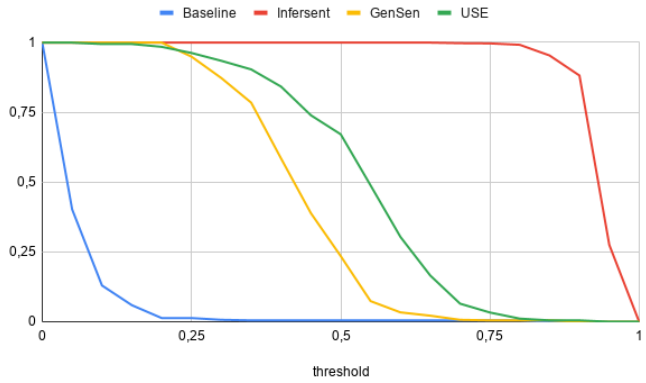

FIGURE 3.4 : Comparison of the average recall by thresholds for all five classes (country, comics character, political party, literary work, film). The threshold takes values from 0.5 to 0.95 by steps of 0.05. The baseline is in blue, InferSent in red, GenSen in yellow and Universal Sentence Encoder in green.

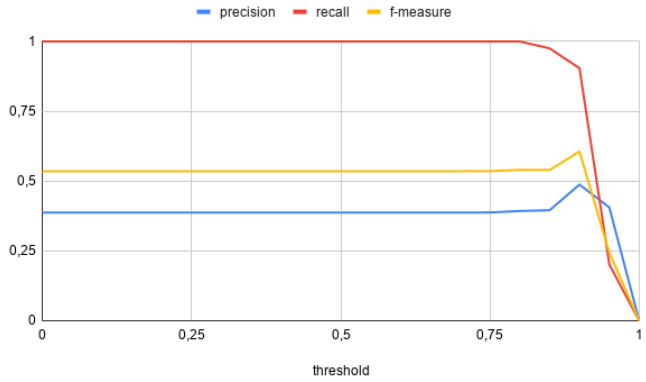

FIGURE 3.5 : Precision (in blue), recall (in red) and F-measure (in yellow) with InferSent and the threshold at 0.9 for the "comics character" class.

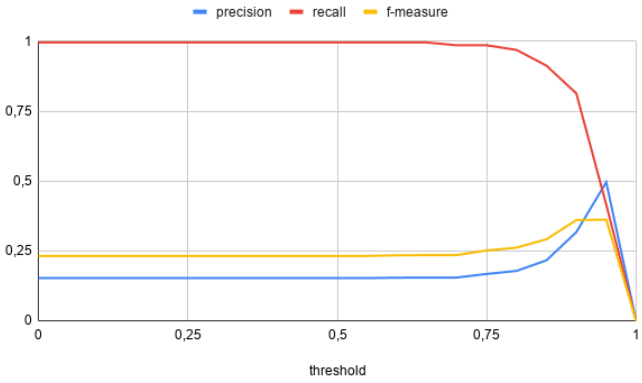

FIGURE 3.6 : Precision (in blue), recall (in red) and F-measure (in yellow) with InferSent and the threshold at 0.9 for the "country" class.

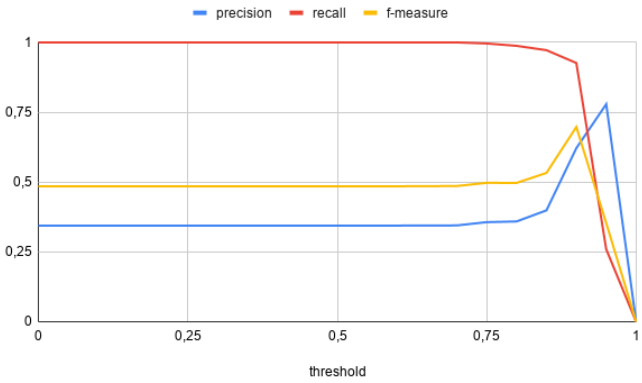

FIGURE 3.7 : Precision (in blue), recall (in red) and F-measure (in yellow) with InferSent and the threshold at 0.9 for the "film" class.

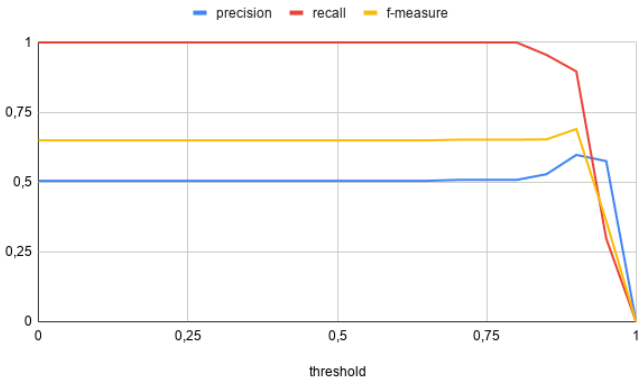

FIGURE 3.8 : Precision (in blue), recall (in red) and F-measure (in yellow) with InferSent and the threshold at 0.9 for the "literary work" class.

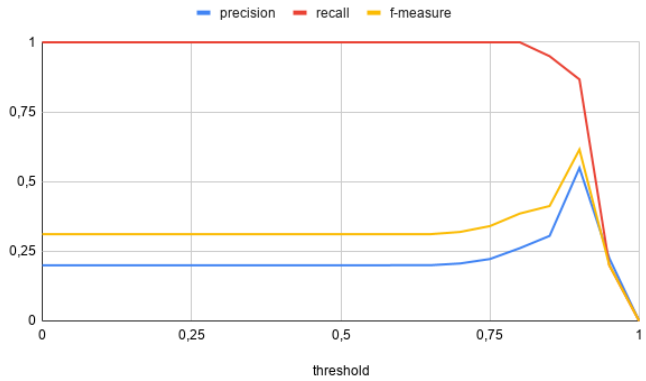

FIGURE 3.9 : Precision (in blue), recall (in red) and F-measure (in yellow) with InferSent and the threshold at 0.9 for the "political party" class.

selected, we obtain thanks to the aforementioned precision formula $3/(3+(12-3)) = 0.25$ as a minimum precision for any possible method of selection. The baseline quickly produces too many false positives without any peak when the threshold is above 0.05. Thus, as expected, the baseline is inadequate. For the three encoders, the same phenomenons append, but with very different threshold values and amplitudes. The three have a peak then a fall. For *GenSen* and *Universal Sentence Encoder* (USE), the peak is rather low and it can be explained by their difficulties to eliminate a sufficient number of false positives. Indeed, property descriptions are relatively close and some of them are well ordered (w.r.t. to their similarity), but not sufficiently to be useful to remove wrong candidates. The peak with *InferSent* is more interesting since it appears later and surpasses 0.5. Hence, *InferSent* can eliminate more false positives than the two others. When looking at the output of the algorithm, right candidates tend to be more grouped at the top of the list, while the two other encoders tend to mix right and wrong candidates. Moreover, because the peaks do not last, it means that the descriptions of the right candidates are very close, and for the three encoders the fall is more or less sudden.

Figure 3.4 shows the evolution of the average recall of the approaches with the same colors as in Figure 3.3. As previously, the same pattern can be observed for all approaches except the baseline, but for very different thresholds. Indeed, immediately the baseline selects too many false negatives, demonstrating once again that it is not suitable to discriminate the right candidates from wrong ones. For the three encoders, at first, all properties are selected, thus there is no false negative. Because descriptions are close in terms of semantics, the cosine similarity produces similarities that are in a close range. Hence, (almost) all properties are detected as right candidates or (almost) none are. Then, very quickly for *GenSen* and USE, there is a sudden drop of the recall. Again, for both of them, the right candidates are distributed among the wrong candidates in a relatively uniform manner. Both encoders are unable to properly sort the right candidates on top of the list and the wrong ones at the bottom. While *InferSent* maintains its good selectivity a lot longer (until a threshold of 0.9). The F-measure of *GenSen* and USE never rises, to the contrary of *InferSent* that reaches 0.59. The latter produces vectors that are closer in their space than the two others, hence the range of cosine similarities is more compacted with *InferSent*. For example, for the entity "Wally West", with *InferSent* the highest similarity score with Π for a candidate property is 0.92 and the worst is 0.75. More important, *InferSent* is able to order more constantly and in a better way the properties. From those results, *InferSent* could be the right candidate to propose to the user an ordered list of candidate properties for propagation.

The second series of results, presented in Figures 3.6 and 3.7, illustrates the behavior of our approach with *InferSent*, since it produces the best results, for each of the five classes. As a reminder, for each class we randomly selected 20 entities, thus, for example, Figure 3.7 shows the average precision in blue of the 20 film entities, the average recall in red and the average F-measure in yellow. Of course, thresholds are the same as in the first series of results. The F-measure is always the better between 0.8 and 0.95, meaning that our approach is extremely sensitive to threshold variations. For all classes, the pattern is the same

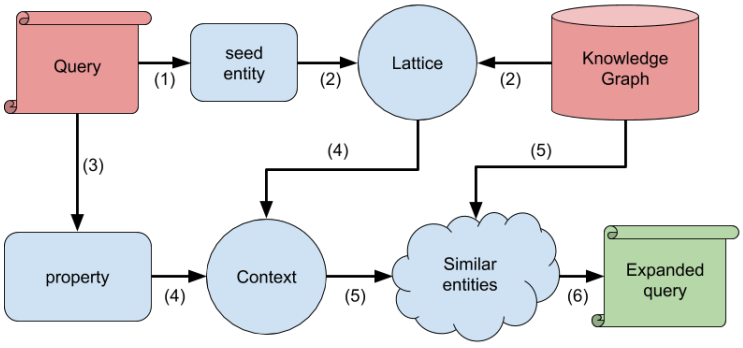

FIGURE 3.10 : Qualitative experiment workflow : the elements in red are the inputs and the element in green is the output. To simplify the diagram, we consider only one instantiated entity linked to one instantiated property in the query.

for the three measures. We only present two out of five classes to keep it concise, all results are available on the GitHub repository. The recall behaves the same for the five classes, but for the precisions, there are more differences. Indeed, comics character and literary work have a quasi-flat part after the peak, meaning that descriptions of right candidates are more distant in the embedded space than the one's of other classes. Hence, similarities of both wrong and right candidates of country, film and political party entities are very close, since, after the peak, the fall is very sudden. Also, the fact that countries and political parties are described with many more properties on average appears here because their precisions start very below other precisions. It is more difficult for our approach to distinguish the right candidates from wrong ones when there are too many candidate properties. For comics characters and literary work entities, the approach seems to be less efficient in removing false positives, maybe because the range of similarity values is too narrow.

Since the choice of propagable properties is subjective for a given set of indiscernibility, three experts may not be sufficient. To establish the gold standard, a crowdsourcing approach might be more appropriate and should merit a formal investigation. As a result, the identity contexts of the gold standard could be more precise. Nevertheless, this approach could be used at least to present to the user an ordered list of candidate properties for propagation, hence helping her to make an educated decision. As a matter of fact, the good recall allows keeping almost all good candidates when the precision may help the user to quickly choose between available properties of the list.

However, we believe that our gold standard is sufficiently well built to state that our results are conclusive. False negatives are due to the fact that some properties are not semantically related to any indiscernible property of the context, and false positives ones are due to some properties that are semantically related to the context but not propagable. Hence, it is obvious that in some cases, considering only the semantic relatedness is a naïve approach. In addition to that, the lack of string-based description on a property is prohibitive, since our approach is based on property description consumption. Taking into account RDF semantics or using other embedding techniques should improve the results.

3.4.3 Qualitative Study

In this section, we introduce three different queries that could benefit from our approach by extending their results. To achieve our goal, we used *InferSent* and a threshold value equal to 0.9. All of these queries are simplified queries tested on Wikidata (for ease of reading). The original queries can be found on the GitHub repository.

Task description

For each query, the goal is to find an identity context that will allow expanding the query with similar entities according to the user's objective. In this way, users can benefit from more complete results. The workflow is the following (see Figure 3.10) : first, from the query, we extract the instantiated entity (or

entities) that will be the seed(s) (step 1). Second, for each seed, we compute its identity lattice (steps 2, 3 and 5). As explained, we build the indiscernibility sets, then we get similar and logically identical entities, then we get candidate properties for propagation and finally, we get propagable properties (see Section 3.3). Third, with the instantiated property (or set of properties) linked to the seed in the query, we select from the lattice, the node having this property in its propagation set (step 7). This node will be considered as the identity context of the query. Indeed, if multiple identity contexts are possible, the user must choose the best suited for its task purpose. Finally, based on the selected identity context, we can rewrite the query with the seed, logically identical entities and similar entities (steps 9 and 11). Furthermore, users can decide to check if the schema of the seed entity is more complete with LOD-CM (see Chapter 5) in steps 8 and 10.

Queries

We tested our approach with three queries. The first query in Listing 3.2 is about the "Paracetamol" drug. The query purpose is to retrieve all clinical trials of this drug. An interesting expansion of this query could be to find all trials of similar legal drugs in terms of medical conditions treated and physical interactions.

```
SELECT DISTINCT ?clinicalTrial WHERE
{
  ?clinicalTrial :researchIntervention :Paracetamol .
}
```

Listing 3.2 : SPARQL query retrieving all studies about the painkiller named Paracetamol.

	Listing 3.2	**Listing 3.4**	**Listing 3.5**
Seed	Paracetamol	France	The Republicans
Ψ	research intervention	head of	member of political party
Π	condition treated, interacts with, legal status	capital, official language	country, political ideology
Similar entities	Ibuprofen Aspirin	French 2nd Republic, July Monarchy, French 3rd Republic, Bourbon Restoration, Kingdom of France, 2nd French Empire, Vichy France, French 4th Republic, 1st French Empire, Paris Commune	UMP, RPR, UDR, UNR UNR
# of results w/o context	586	12	2
# of results w/ context	860	99 (77)	13

TABLE 3.1 : Identity context contribution to queries.

Table 3.1 shows additional results brought by our approach. Each column corresponds to a query. For the first query (and also for the next ones), there is only one seed "Paracetamol" ("France" and "the Republicans" in the second and the third columns respectively)as it is the only instantiated entity in the

query. To fill this table, we first computed the lattice of the seed, then, selected a context containing the property "research intervention" in its Ψ. We chose as a context, legal drugs having the same medical conditions and the same physical interactions (obviously, any other context could be chosen depending on the users' needs). Finally, the query is expanded with similar entities as shown in Listing 3.3. The results show a 47% increase in the number of clinical trials for the considered context.

```
SELECT DISTINCT ?clinicalTrial WHERE
{
  VALUES (?drug)
  {
    (:Paracetamol) (:Ibuprofen) (:Aspirin)
  }
  ?clinicalTrial :researchIntervention ?drug .
}
```

Listing 3.3 : Expanded SPARQL query retrieving all studies about Paracetamol similar entities.

The second query, in Listing 3.4, is about retrieving persons who once lead France. However, France has a complex history and has changed its political regime several times (for example, during World War II, or during the Napoleonian period). Thus, even if the French territory was "almost" always the same during the past centuries, each political regime has its own entity in Wikidata. Therefore, the query might not give all expected results. But if the user chooses the right identity context, i.e., $\mathcal{C}_{(\{capital,officialLanguage\},\{headOf\},\approx)}$ then all expected people will be retrieved.

```
SELECT DISTINCT ?headOfState WHERE
{
  ?headOfState :headOf :France .
}
```

Listing 3.4 : SPARQL query retrieving all people who were head of the French state.

Similarly to the query about "Paracetamol", we computed the lattice and search for context with *headOf* in the propagable properties. The results are shown in the second column of Table 3.1. The expanded query could be rewritten as the previous one. It should be noted that among the 99 results, 22 persons were not head of France. Fourteen were head of Paris City Council and 8 were Grand Master of a Masonic obedience in France. This is due to the fact that the council and the obedience are misplaced in the Wikidata ontology. These errors cannot, therefore, be attributed to our approach. The results show a 542% increase in the number of France leaders for the considered context.

Finally, in Listing 3.5, we present a query about French politicians from The Republican party that have been convicted. The peculiarity here is that this major political party changed its name several times because of either political scandal or humiliating defeats. Consequently, if the Knowledge Graph is not up to date or not complete, some persons who were members of multiple versions of this party in the real world could not be actually linked to each version in the Knowledge Graph. This is the case of Wikidata that returns, for the query of Listing 3.5, only two politicians. However, there are more than a dozen politicians of this party who have been convicted of various felonies. By using our approach, it is possible to select a context composed of the political alignment and the country for which the *memberOf* property is propagable, and, hence, obtain a more complete result (of course depending on the completeness of data about politicians in Wikidata).

```
SELECT DISTINCT ?politician ?crime WHERE
{
  ?politician :memberOf :TheRepublicans ;
    :convictedOf ?crime .
}
```

Listing 3.5 : SPARQL query retrieving all politicians member of French party named The Republicans that were convicted.

The same steps as for queries about "Paracetamol" and "France" were reproduced. Results are shown in the third column of Table 3.1. The results show a 550% increase in the number of convicted politicians for the considered context.

3.4.4 Discussion

As we have seen, our approach allows discovering propagable properties for a given indiscernibility set of properties Π. An identity context with its indiscernibility and propagation sets can provide more complete answers to queries through query expansion. The results are very promising but need to be confronted with more different kinds of knowledge graphs and a combination of distinct knowledge graphs. Also, our approach does not work well when the property of an entity lack property describing it (such as *rdfs:comment* or *schema:description*). It is a limitation since some ontologies do not provide textual descriptions of their properties. Hence, the first step for future work is to circumvent this flaw with a multi-faceted approach. Moreover, in a textual description, some words might be irrelevant (like a Wikidata identifier) and degrade the quality of the results.

3.5 Conclusion

In this chapter, we demonstrated that propagating properties can be discovered semi-automatically. To this end, we presented an approach based on sentence embedding. Given an indiscernible set of properties, the proposed system discovers properties that could be propagated using semantic similarities between the properties. Our approach computes, for an entity, an identity lattice that represents all its possible identity contexts, i.e., both indiscernible and propagating properties. We validated using quantitative and qualitative evaluations that the proposed approach generates promising results for both discovering propagating properties and providing complete answers to the given queries.

Future work includes using other features to improve the results, like values of properties, number of property usage, or semantic features of the property should be tried. However, capturing ontological information of a property when embedding is still an open problem. Secondly, using only sentence embedding, combined with intuition from Tober's first law, might be naïve in some cases. Therefore, there is a need to challenge our work with a combination of distinct KGs. For the time being, we only considered in lattices the case where the entity is subject to a triple, and we should also consider cases where it is the value of a triple. Moreover, using SPARQL queries to help the user to select the best-suited identity context might be an interesting starting point for later work. Finally, to explore SPARQL queries expansion (presented in Section 3.4.3), a prototype should be implemented to allow users selecting the proper context according to an ordered list of contexts. Also, using RDF* and/or SPARQL* [HT14] to represent the context as defined in this chapter should be investigated.

Enriching KGs with Fuzzy Temporal Data

This chapter is based on work realized in the context of the VIVA project and the PhD theses of Fatma Ghorbel (co-supervised with Elisabeth Métais, Nebrasse Ellouze, and Faïez Gargouri) and Noura Herradi (co-supervised with Elisabeth Métai). [GHM20], [GHM+19b], [GHM19a], [MGH+18]

Enriching Knowledge Graphs with temporal information could be a very complex process. Indeed, temporal information given by users is often imprecise. For instance, if they give the information "Alexandre was married to Nicole by 1981 to late 90" two measures of imprecision are involved. On the one hand, the information "by 1981" is imprecise in the sense that it could mean approximately from 1980 to 1982 ; on the other hand, the information "late 90" is imprecise in the sense that it could mean, with an increasingly possibility, from 1995 to 2000. When an event is characterized by a gradual beginning and/or ending, it is usual to represent the corresponding time span as an imprecise time interval.

In the Semantic Web field, several approaches have been proposed to represent and reason about precise temporal data. However, most of them handle only precise time intervals and associated qualitative relations i.e., they are not intended to handle time points and qualitative relations between a time interval and a time point or two time points. Besides, to the best of our knowledge, there is no approach devoted to handle imprecise temporal data.

In this chapter, I present the two approches that we proposed to address this problem :

— The first approach involves only crisp [1] standards and tools. To represent imprecise time intervals in OWL 2, we extend the so called 4D-fluents model [WF06] which is a formalism to model crisp quantitative temporal information and the evolution of temporal concepts in OWL. This model is extended in two ways : *(i)* it is enhanced with new crisp components for modeling imprecise time intervals, and, *(ii)* with qualitative temporal expressions representing crisp relations between imprecise temporal intervals. To reason on imprecise time intervals, we extend the Allen's interval algebra [All83] which is the most used and known formalisms for reasoning about crisp time intervals. We generalize Allen's relationships to handle imprecise time intervals with a crisp view. The resulting crisp temporal interval relations are inferred from the introduced imprecise time intervals using a set of SWRL rules [HPSB+04], in OWL 2.

— The second approach is based on fuzzy sets theory and fuzzy tools. It is based on Fuzzy-OWL 2 [BS11] which is an extension of OWL 2 that deals with fuzzy information. To represent imprecise time intervals in Fuzzy-OWL 2, we extend, as for crisp, the 4D-fluents model in two ways : *(i)* with new fuzzy components to be able to model imprecise time intervals, and, *(ii)* with qualitative temporal expressions representing fuzzy relations between imprecise temporal intervals. To reason on imprecise time intervals, we extend Allen's work to compare imprecise time intervals in a fuzzy gradual personalized way. Our Allen's extension introduces gradual fuzzy interval relations e.g., "long before". It is personalized in the sense that it is not limited to a given number of interval relations. It is possible to determinate the level of precision that should be in a given context. For instance, the classic Allen relation "before" may be generalized in N interval relations, where "before(1)" means "just before" and gradually the time gap between the two imprecise intervals increases until "before(N)" which means "long before". The resulting fuzzy interval relations are inferred from the introduced imprecise time intervals using the FuzzyDL reasoner [BS08], via a set of Mamdani IF-THEN rules, in Fuzzy-OWL 2.

The current chapter is organized as follows : Section 4.1 is devoted to present some preliminary concepts and related work in the field of temporal information representation in OWL and reasoning on time intervals. In Section 4.2, we introduce our crisp-based approach for representing and reasoning on imprecise

1. The word "crisp" designates "precise", in opposite to "fuzzy" in the context of fuzzy sets theory. An ontology is either "crisp" (i.e., a "classic" ontology) or "fuzzy".

TABLE 4.1 : Allen's temporal interval relations (I : ——— , J : ═══).

Relation	Inverse	Relations between interval bounds	Illustration
$Before(I,J)$	$After(I,J)$	$I^+ < J^-$	
$Meets(I,J)$	$MetBy(I,J)$	$I^+ = J^-$	
$Overlaps(I,J)$	$OverlappedBy(I,J)$	$(I^- < J^-) \wedge (I^+ > J^-) \wedge (I^+ < J^+)$	
$Starts(I,J)$	$StartedBy(I,J)$	$(I^- = J^-) \wedge (I^+ < J^+)$	
$During(I,J)$	$Contains(I,J)$	$(I^- > J^-) \wedge (I^+ < J^+)$	
$Ends(I,J)$	$EndedBy(I,J)$	$(I^- > J^-) \wedge (I^+ = J^+)$	
$Equal(I,J)$	$Equal(I,J)$	$(I^- = J^-) \wedge (I^+ = J^+)$	

time intervals. In Section 4.3, we introduce our fuzzy-based approach for representing and reasoning on imprecise time intervals. Section 4.4 describes the application of our approches in two different fields and draws conclusions.

4.1 Preliminaries and State of the Art

In this section, we introduce some preliminary concepts and related work in the field of temporal information representation in OWL and reasoning on time intervals.

4.1.1 Representing Temporal Information in OWL

Five main approaches are proposed to represent time information in OWL : Temporal Description Logics [AF00], Versioning [KF01], N-ary relations [NR06] and 4D-fluents [WF06]. All these approaches represent only crisp temporal information in OWL. Temporal Description Logics extend the standard description logics with additional temporal constructs e.g., "sometime in the future". N-ary relations approach represents an N-ary relation using an additional object. The N-ary relation is represented as two properties each related with the new object. The two objects are related to each other with an N-ary relation. Reification is "a general purpose technique for representing N-ary relations using a language such as OWL that permits only binary relations" [BP11]. Versioning approach is described as "the ability to handle changes in ontologies by creating and managing different variants of it" [KF01]. When an ontology is modified, a new version is created to represent the temporal evolution of the ontology. 4D-fluents approach represents temporal information and the evolution of the last ones in OWL. Concepts varying in time are represented as 4-dimensional objects with the 4th dimension being the temporal dimension.

Based on the present related work, we choose the 4D-fluents approach. Indeed, compared to related work, it minimizes the problem of data redundancy as the changes occur only on the temporal parts and keeping therefore the static part unchanged. It also maintains full OWL expressiveness and reasoning support [BP11]. We extend this approach in two ways. (1) It is extended with crisp components to represent imprecise time intervals and crisp interval relations in OWL 2 (Section 4.2). (2) It is extended with fuzzy components to represent imprecise time intervals and fuzzy interval relations in Fuzzy-OWL 2 (Section 4.3).

4.1.2 Allen's Interval Algebra

[All83] has proposed 13 mutually exclusive primitive relations that may hold between two precise time intervals. Their semantics is illustrated in Table 4.1. Let $I = [I^-, I^+]$ and $J = [J^-, J^+]$ two time intervals ; where I^- (respectively J^-) is the beginning time-step of the event and I^+ (respectively J^-) is the ending.

A number of works fuzzify Allen's temporal interval relations. We classify these works into (1) works focusing on fuzzifying Allen's interval algebra to compare precise time intervals and (2) works focusing on fuzzifying Allen's interval algebra to compare imprecise time intervals.

Three approaches have been proposed to fuzzify Allen's interval algebra in order to compare precise time intervals : [GHP94], [DP89] and [BG06]. [GHP94] propose fuzzy Allen relations viewed as fuzzy sets of ordinary Allen relationship taking into account a neighborhood structure, a notion originally introduced in

[Fre92]. [DP89] represent a time interval as a pair of possibility distributions that define the possible values of the endpoints of the crisp interval. Using possibility theory, the possibility and necessity of each of the interval relations can then be calculated. This approach also allows modeling imprecise relations such as "long before". [BG06] propose a fuzzy extension of Allen's work, called IAfuz where degrees of preference are associated to each relation between two precise time intervals.

Four approaches have been proposed to fuzzify Allen's interval algebra to compare imprecise time intervals : [NM03], [Ohl04], Schockaert08 and [GHY17]. [NM03] propose a temporal model based on fuzzy sets to extend Allen relations with imprecise time intervals. The authors introduce a set of auxiliary operators on intervals and define fuzzy counterparts of these operators. The compositions of these relations are not studied by the authors. [Ohl04] propose an approach to handle some gradual temporal relations as "more or less finishes". However, this work cannot take into account gradual temporal relations such as "long before". Furthermore, many of the symmetry, reflexivity, and transitivity properties of the original temporal interval relations are lost in this approach ; thus it is not suitable for temporal reasoning. [SC08] propose a generalization of Allen's relations with precise and imprecise time intervals. This approach allows handling classical temporal relations, as well as other imprecise relations. Interval relations are defined according to two fuzzy operators comparing two time instants : "long before" and "occurs before or at approximately the same time". [GHY17] generalize the definitions of the 13 Allen's classic interval relations to make them applicable to fuzzy intervals in two ways (conjunctive and disjunctive). Gradual temporal interval relations are not taken into account.

4.2 A Crisp-Based Approach for Representing and Reasoning on Imprecise Time Intervals

In this section, we propose a crisp-based approach to represent and reason on imprecise time intervals. This solution is entirely based on crisp standards and tools. We extend the 4D-fluents model to represent imprecise time intervals and their crisp relationships in OWL 2. To reason on imprecise time intervals, we extend the Allen's interval algebra in a crisp way. In OWL 2, inferences are done via a set of SWRL rules.

4.2.1 Representing Imprecise Time Intervals and Crisp Qualitative Interval Relations in OWL 2

In the crisp-based solution, we now represent each imprecise interval bound of the time interval as a disjunctive ascending set. Let $I = [I^-, I^+]$ be an imprecise time interval ; where $I^- = I^{-(1)} \ldots I^{-(N)}$ and $I^+ = I^{+(1)} \ldots I^{+(N)}$. For instance, if we have the information "Alexandre was started his PhD study in 1975 and he was graduated around 1980" the imprecise time interval representing this period is $[1975, \{1978 \ldots 1982\}]$. This means that his PhD studies end in 1978 or 1979 or 1980 or 1981 or 1982. The classic 4D-fluents model introduces two crisp classes "TimeSlice" and "TimeInterval" and four crisp properties "tsTimeSliceOf", "tsTimeInterval", "hasBegining" and "hasEnd". The class "TimeSlice" is the domain class for entities representing temporal parts (i.e., "time slices"). The property "tsTimeSliceOf" connects an instance of class "TimeSlice" with an entity. The property "tsTimeInterval" connects an instance of class "TimeSlice" with an instance of class "TimeInterval". The instance of class "TimeInterval" is related with two temporal instants that specify its starting and ending points using, respectively, the "hasBegining" and "hasEnd" properties. Figure 4.1 illustrates the use of the 4D-fluents model to represent the following example : "Alexandre was started his PhD study in 1975 and he was graduated in 1978".

We extend the original 4D-fluents model in the following way. We add four crisp datatype properties "HasBeginningFrom", "HasBeginningTo", "HasEndFrom", and "HasEndTo" to the class "TimeInterval". Let $I = [I^-, I^+]$ be an imprecise time interval ; where $I^- = I^{-(1)} \ldots I^{-(N)}$ and $I^+ = I^{+(1)} \ldots I^{+(N)}$. "HasBeginningFrom" has the range $I^{-(1)}$. "HasBeginningTo" has the range $I^{-(N)}$. "HasEndFrom" has the range $I^{+(1)}$. "HasEndTo" has the range $I^{+(N)}$. The 4D-fluents model is also enhanced with crisp qualitative temporal interval relations that may hold between imprecise time intervals. This is implemented by introducing temporal relationships, called "RelationIntervals", as a crisp object property between two instances of the class "TimeInterval". Figure 4.2 represents the extended 4D-fluents model in OWL 2.

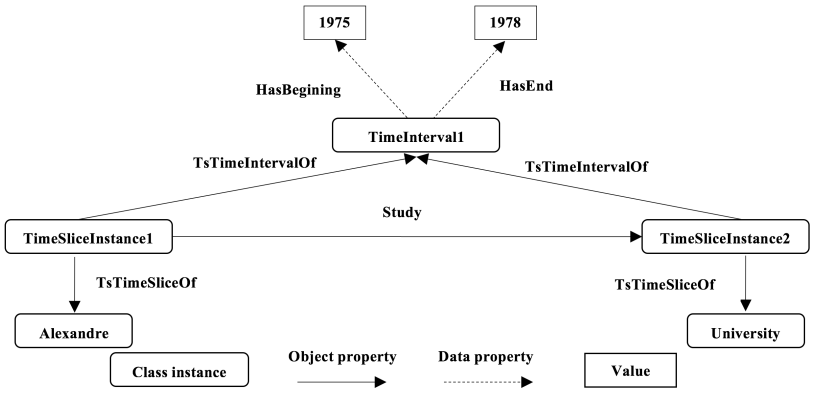

FIGURE 4.1 : An instantiation of the classic the 4D-fluents model.

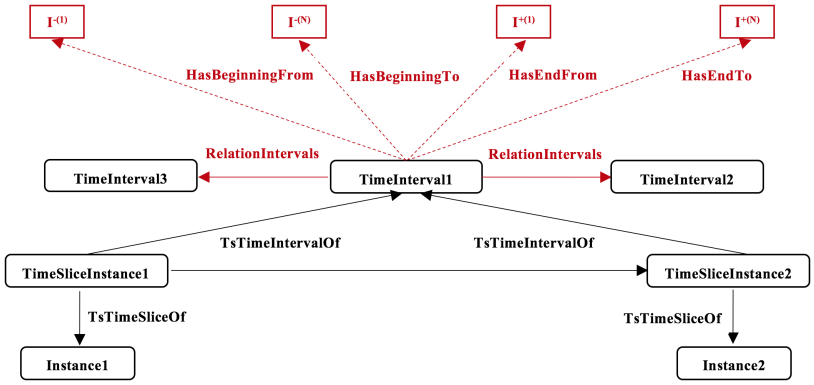

FIGURE 4.2 : The extended 4D-fluents model in OWL 2.

We can see in Figure 4.3 an instantiation of the extended 4D-fluents model in OWL 2. On this example, we consider the following information : "Alexandre was married to Nicole just after he was graduated with a PhD. Alexandre was graduated with a PhD in 1980. Their marriage lasts 15 years. Alexandre was remarried to Béatrice since about 10 years and they were divorced in 2016". Let $I = [I^-, I^+]$ and $J = [J^-, J^+]$ be two imprecise time intervals representing, respectively, the duration of the marriage of Alexandre with Nicole and the one with Béatrice. Assume that $I^- = \{1980 \ldots 1983\}$, $I^+ = \{1995 \ldots 1998\}$, $J^- = \{2006 \ldots 2008\}$ and $J^+ = 2016$.

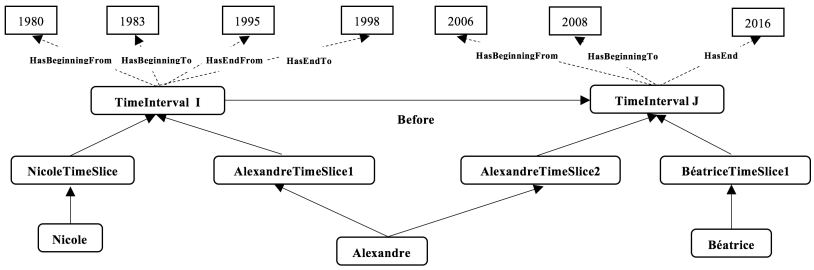

FIGURE 4.3 : An instantiation of the extended 4D-fluents model in OWL 2.

4.2.2 A Crisp-Based Reasoning on Imprecise Time Intervals in OWL 2

We have redefined 13 Allen's interval, in a crisp way, to compare imprecise time intervals i.e., the resulting interval relations upon imprecise time intervals are crisp. Let $I = [I^-, I^+]$ and $J = [J^-, J^+]$ two imprecise time intervals ; where $I^- = I^{-(1)} \ldots I^{-(N)}$, $I^+ = I^{+(1)} \ldots I^{+(N)}$, $J^- = J^{-(1)} \ldots J^{-(N)}$ and $J^+ = J^{+(1)} \ldots J^{+(N)}$. For instance, the crisp interval relation "before (I, J)" is redefined as : $\forall I^{+(i)} \in I^+, \forall J^{-(j)} \in J^- : I^{+(i)} < J^{-(j)}$ This means that the most recent time instant of $I^+(I^{+(N)})$ ought to Precede the oldest time instant of $J^-(J^{-(1)}) : I^{+(N)} < J^{-(1)}$ In the similar way, we define the other temporal interval relations. In Table 4.2, we define 13 crisp temporal

TABLE 4.2 : Crisp temporal interval relations upon imprecise time intervals.

Relation	Inverse	Interpretation	Relations between interval bounds
$Before(I,J)$	$After(I,J)$	$\forall I^{+(i)} \in I^+, \forall J^{-(j)} \in J^- : (I^{+(i)} < J^{-(j)})$	$I^{+(N)} < J^{-(1)}$
$Meets(I,J)$	$MetBy(I,J)$	$\forall I^{+(i)} \in I^+, \forall J^{-(j)} \in J^- : (I^{+(i)} = J^{-(j)})$	$(I^{+(1)} = J^{-(1)}) \wedge (I^{+(N)} = J^{-(N)})$
$Overlaps(I,J)$	$OverlappedBy(I,J)$	$\forall I^{-(i)} \in I^-, \forall I^{+(i)} \in I^+, \forall J^{-(j)} \in J^-, \forall J^{+(j)} \in J^+ : (I^{-(i)} < J^{-(j)}) \wedge (J^{-(j)} < I^{+(i)}) \wedge (I^{+(i)} < J^{+(j)})$	$(I^{-(N)} < J^{-(1)}) \wedge (J^{-(N)} < I^{+(1)}) \wedge (I^{+(N)} < J^{+(1)})$
$Starts(I,J)$	$StartedBy(I,J)$	$\forall I^{-(i)} \in I^-, \forall I^{+(i)} \in I^+, \forall J^{-(j)} \in J^-, \forall J^{+(j)} \in J^+ : (I^{-(i)} = J^{-(j)}) \wedge (I^{+(i)} < J^{+(j)})$	$(I^{-(1)} = J^{-(1)}) \wedge (I^{-(N)} = J^{-(N)}) \wedge (I^{+(N)} < J^{+(1)})$
$During(I,J)$	$Contains(I,J)$	$\forall I^{-(i)} \in I^-, \forall I^{+(i)} \in I^+, \forall J^{-(j)} \in J^-, \forall J^{+(i)} \in J^+ : (J^{-(j)} < I^{-(i)}) \wedge (I^{+(i)} < J^{+(i)})$	$(J^{-(N)} < I^{-(1)}) \wedge (I^{+(N)} < J^{+(1)})$
$Ends(I,J)$	$EndedBy(I,J)$	$\forall I^{-(i)} \in I^-, \forall I^{+(i)} \in I^+, \forall J^{-(j)} \in J^-, \forall J^{+(j)} \in J^+ : (I^{-(i)} < J^{-(j)}) \wedge (I^{+(i)} = J^{+(j)})$	$(J^{-(N)} < I^{-(1)}) \wedge (I^{+(1)} = J^{+(1)}) \wedge (I^{+(N)} = J^{+(N)})$
$Equal(I,J)$	$Equal(I,J)$	$\forall I^{-(i)} \in I^-, \forall I^{+(i)} \in I^+, \forall J^{-(j)} \in J^-, \forall J^{+(j)} \in J^+ : (I^{-(i)} = J^{-(j)}) \wedge (I^{+(i)} = J^{+(j)})$	$(I^{-(1)} = J^{-(1)}) \wedge (I^{-(N)} = J^{-(N)}) \wedge (I^{+(1)} = J^{+(1)}) \wedge (I^{+(N)} = J^{+(N)})$

interval relations upon the two imprecise time intervals I and J.

In order to apply our crisp extension of Allen's work in OWL 2, we propose a set of SWRL rules that infer the temporal interval relations from the introduced imprecise time intervals which are represented using the extended 4D-fluents model in OWL2. For each temporal interval relation, we associate a SWRL rule. Reasoners that support DL-safe rules (i.e., rules that apply only on named individuals in the knowledge base) such as Pellet [SPG+07] can support our approach. For instance, the SWRL rule to infer the "Meet (I, J)" relation is the following :

$$TimeInterval(I) \wedge TimeInterval(J) \wedge HasEndFrom(I,a) \wedge HasBeginningFrom(J,b) \wedge Equals(a,b) \wedge HasEndTo(I,c) \wedge HasBeginningTo(J,d) \wedge Equals(c,d) \rightarrow Meet(I,J)$$

4.3 A Fuzzy-Based Approach for Representing and Reasoning on Imprecise Time Intervals

In this section, we propose a fuzzy-based approach to represent and reason on imprecise time intervals. This approach is based on a fuzzy environment. We extend the 4D-fluents model to represent imprecise time intervals and their relationships in Fuzzy-OWL 2. To reason on imprecise time intervals, we extend the Allen's interval algebra in a fuzzy gradual personalized way. We infer the resulting fuzzy interval relations in Fuzzy-OWL 2 using a set of Mamdani IF-THEN rules.

4.3.1 Representing Imprecise Time Intervals and Fuzzy Qualitative Interval Relations in Fuzzy-OWL 2

In the fuzzy-based solution, we now represent the imprecise beginning interval bound as a fuzzy set which has the L-function MF and the ending interval bound as a fuzzy set which has the R-function membership function (MF). Let $I = [I^-, I^+]$ be an imprecise time interval. We represent the binging bound I^- as a fuzzy set which has the L-function MF ($A = I^{-(1)}$ and $B = I^{-(N)}$). We represent the ending bound I^+ as a fuzzy set which has the R-function MF ($A = I^{+(1)}$ and $B = I^{+(N)}$). For instance, if we have the information "Alexandre was starting his PhD study in 1973 and was graduated in late 80", the beginning bound is crisp. The ending bound is imprecise and it is represented by L-function MF ($A = 1976$ and $B = 1980$). For the rest of the chapter, we use the MFs shown in Figure 4.4 [Zadeh, 1975].

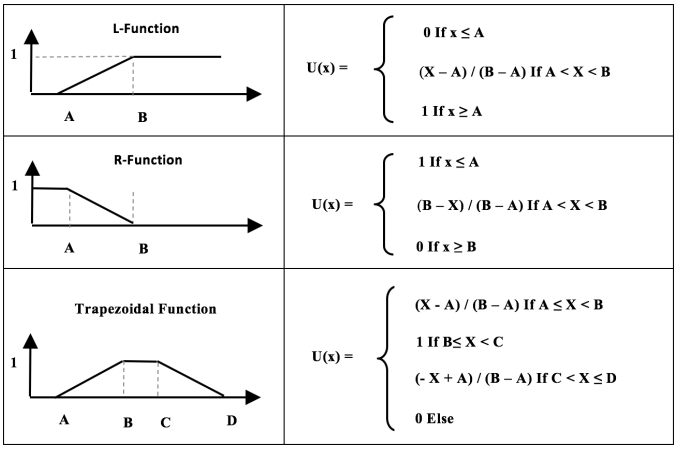

FIGURE 4.4 : R-Function, L-Function and Trapezoidal MFs [Zad75].

We extend the original 4D-fluents model to represent imprecise time intervals in the following way. We add two fuzzy datatype properties "FuzzyHasBegining" and "FuzzyHasEnd" to the class "TimeInterval". "FuzzyHasBegining" has the L-function MF ($A = I^{-(1)}$ and $B = I^{-(N)}$). "FuzzyHasEnd" has the R-function MF ($A = I^{+(1)}$ and $B = I^{+(N)}$). The 4D-fluents approach is also enhanced with qualitative temporal relations that may hold between imprecise time intervals. We introduce the "FuzzyRelationIntervals", as a fuzzy object property between two instances of the class "TimeInterval". "FuzzyRelationIntervals" represent fuzzy qualitative temporal relations. "FuzzyRelationIntervals" has the L-function MF ($A = 0$ and $B = 1$). Figure 4.5 represents our extended 4D-fluents model in Fuzzy-OWL 2.

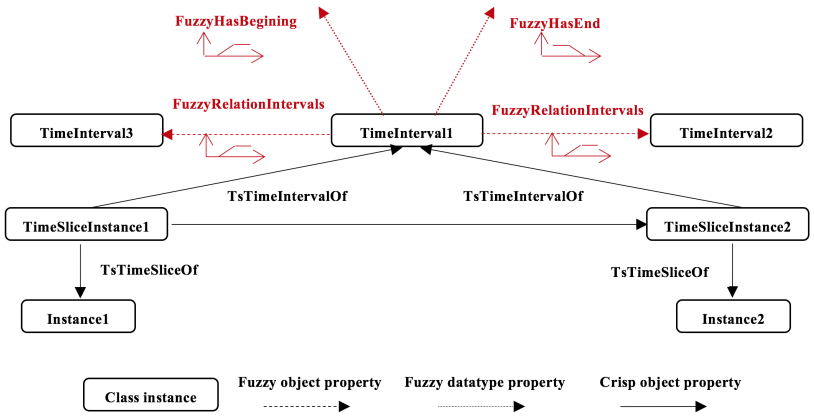

FIGURE 4.5 : The extended 4D-fluents model in Fuzzy-OWL 2.

We can see in Figure 4.6 an instantiation of the extended 4D-fluents model in Fuzzy-OWL 2. On this example, we consider the following information : "Alexandre was married to Nicole just after he was graduated with a PhD. Alexandre was graduated with a PhD in 1980. Their marriage lasts 15 years. Alexandre was remarried to Béatrice since about 10 years and they were divorced in 2016". Let $I = [I^-, I^+]$ and $J = [J^-, J^+]$ be two imprecise time intervals representing, respectively, the duration of the marriage of Alexandre with Nicole and the one with Béatrice. I^- is represented with the fuzzy datatype property "FuzzyHasBegining" which has the L-function MF ($A = 1980$ and $B = 1983$). I^+ is represented with the fuzzy datatype property "FuzzyHasEnd" which has the R-function MF ($A = 1995$ and $B = 1998$). J^- is represented with the fuzzy datatype property "FuzzyHasBegining" which has the L-function MF ($A = 2005$ and $B = 2007$). J^+ is represented with the crisp datatype property "HasEnd" which has the value "2016".

4.3.2 A Fuzzy-Based Reasoning on Imprecise Time Intervals in Fuzzy OWL 2

We propose a set of fuzzy gradual personalized comparators that may hold between two time instants. Based on these operators, we present our fuzzy gradual personalized extension of Allen's work. Then, we

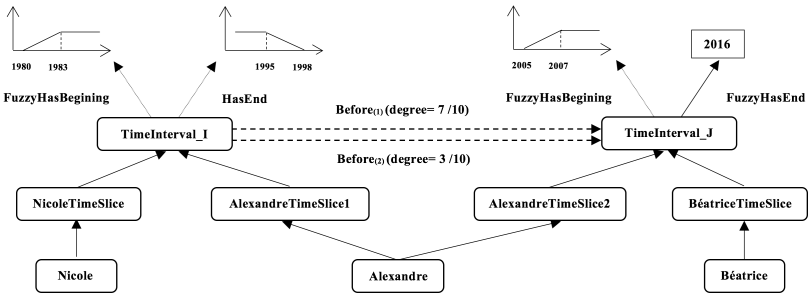

FIGURE 4.6 : An instantiation of the extended 4D-fluents model in Fuzzy-OWL 2.

infer, in Fuzzy OWL 2, the resulting temporal interval relations via a set of Mamdani IF-THEN rules using the fuzzy reasoner FuzzyDL. We generalize the crisp time instants comparators "Follow", "Precede" and "Same", introduced in [VK86]. Let α and β two parameters allowing the definition of the membership function of the following comparators $(\in]0,+\infty[)$; N is the number of slices ; T_1 and T_2 are two time instants ; we define the following comparators (illustrated in Figure 4.7) :

— $\{Follow_{(1)}^{(\alpha,\beta)}(T_1,T_2)\ldots Follow_{(N)}^{(\alpha,\beta)}(T_1,T_2)\}$ are a generalization of the crisp time instants relation "Follows". $Follow_{(1)}^{(\alpha,\beta)}(T_1,T_2)$ means that T_1 is "just after or approximately at the same time" T_2 w.r.t. (α,β) and gradually the time gap between T_1 and T_2 increases until $Follow_{(N)}^{(\alpha,\beta)}(T_1,T_2)$ which means that T_1 is "long after" T_2 w.r.t. (α,β). N is set by the expert domain. $\{Follow_{(1)}^{(\alpha,\beta)}(T_1,T_2)\ldots Follow_{(N)}^{(\alpha,\beta)}(T_1,T_2)\}$ are defined as fuzzy sets. $Follow_{(1)}^{(\alpha,\beta)}(T_1,T_2)$ has R-Function MF which has as parameters $A=\alpha$ and $B=(\alpha+\beta)$. All comparators $\{Follow_{(2)}^{(\alpha,\beta)}(T_1,T_2)\ldots Follow_{(N-1)}^{(\alpha,\beta)}(T_1,T_2)\}$ have trapezoidal MF which has as parameters $A=((K-1)\alpha)$ and $B=((K-1)\alpha+(K-1)\beta)$, $C=(K\alpha+(K-1)\beta)$ and $D=(K\alpha+K\beta)$; where $2\leq K\leq N-1$. $Follow_{(N)}^{(\alpha,\beta)}(T_1,T_2)$ has L-Function MF which has as parameters $A=((N-1)\alpha+(N-1)\beta)$ and $B=((N-1)\alpha+(N-1)\beta)$;

— $\{Precede_{(1)}^{(\alpha,\beta)}(T_1,T_2)\ldots Precede_{(N)}^{(\alpha,\beta)}(T_1,T_2)\}$ are a generalization of the crisp time instants relation "Precede". $Precede_{(1)}^{(\alpha,\beta)}(T_1,T_2)$ means that T_1 is "just before or approximately at the same time" T_2 w.r.t. (α,β) and gradually the time gap between T_1 and T_2 increases until $Precede_{(N)}^{(\alpha,\beta)}(T_1,T_2)$ which means that T_1 is "long before" T_2 w.r.t. (α,β). N is set by the expert domain. $\{Precede_{(1)}^{(\alpha,\beta)}(T_1,T_2)\ldots Precede_{(N)}^{(\alpha,\beta)}(T_1,T_2)\}$ are defined as fuzzy sets. $Precede_{(i)}^{(\alpha,\beta)}(T_1,T_2)$ is defined as :

$$Precede_{(i)}^{(\alpha,\beta)}(T_1,T_2)=1-Follow_{(i)}^{(\alpha,\beta)}(T_1,T_2)$$

— We define the comparator $Same^{(\alpha,\beta)}$ which is a generalization of the crisp time instants relation "Same". $Same^{(\alpha,\beta)}(T_1,T_2)$ means that T_1 is "approximately at the same time" T_2 w.r.t. (α,β). It is defined as :

$$Same^{(\alpha,\beta)}(T_1,T_2)=Min(Follow_{(1)}^{(\alpha,\beta)}(T_1,T_2),\ Precede_{(1)}^{(\alpha,\beta)}(T_1,T_2))$$

Then, we extend Allen's work to compare imprecise time intervals with a fuzzy gradual personalized view. We provide a way to model gradual, linguistic-like description of temporal interval relations. Compared to related work, our work is not limited to a given number of imprecise relations. It is possible to determinate the level of precision that should be in a given context. For instance, the classic Allen relation "before" may be generalized in N imprecise relations, where "$Before_{(1)}^{(\alpha,\beta)}(I,J)$" means that I is "just before" J w.r.t. (α,β) and gradually the time gap between I and J increases until "$Before_{(N)}^{(\alpha,\beta)}(I,J)$" which means that I is long

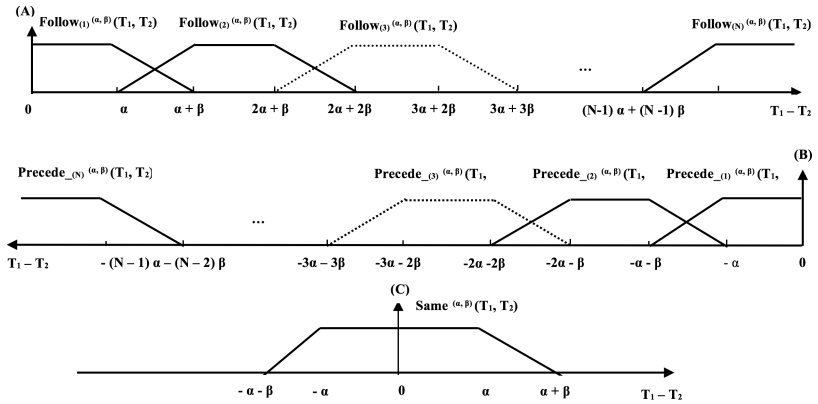

FIGURE 4.7 : Fuzzy gradual personalized time instants comparators (A) Fuzzy sets of $\{Follow^{(\alpha,\beta)}_{(1)}(T_1,T_2)\ldots Follow^{(\alpha,\beta)}_{(N)}(T_1,T_2)\}$. (B) Fuzzy sets of $\{Precede^{(\alpha,\beta)}_{(1)}(T_1,T_2)\ldots Precede^{(\alpha,\beta)}_{(N)}(T_1,T_2)\}$. (C) Fuzzy set of $Same^{(\alpha,\beta)}(T_1,T_2)$.

before J w.r.t. (α,β). The definition of our fuzzy interval relations is based on the fuzzy gradual personalized time instants compactors. Let $I=[I^-,I^+]$ and $J=[J^-,J^+]$ two imprecise time intervals; where I^- has the L-function MF ($A=I^{-(1)}$ and $B=I^{-(N)}$); I^+ is a fuzzy set which has the R-function MF ($A=I^{+(1)}$ and $B=I^{+(N)}$); J^- is a fuzzy set which has the L-function MF ($A=J^{-(1)}$ and $B=J^{-(N)}$); J^+ is a fuzzy set which has the R-function MF ($A=J^{+(1)}$ and $B=J^{+(N)}$). For instance, the fuzzy interval relation "$Before^{(\alpha,\beta)}_{(1)}(I,J)$" is defined as :

$$\forall I^{+(i)}\in I^+,\forall J^{-(j)}\in J^-:$$
$$Precede^{(\alpha,\beta)}_{(1)}(I^{+(i)},J^{-(j)})$$

This means that the most recent time instant of $I^+(I^{+(N)})$ ought to proceed the oldest time instant of $J^-(J^{-(1)})$:

$$Precede^{(\alpha,\beta)}_{(1)}(I{+}(N),J{-}(1))$$

In the similar way, we define the others temporal interval relations, as shown in Table 4.3.

Finally, we have implemented our fuzzy gradual personalized extension of Allen's work in Fuzzy-OWL 2. We use the ontology editor PROTEGE version 4.3 and the fuzzy reasoner FuzzyDL. We propose a set of Mamdani IF-THEN rules to infer the temporal interval relations from the introduced imprecise time intervals which are represented using the extended 4D-fluents model in Fuzzy-OWL2. For each temporal interval relation, we associate a Mamdani IF-THEN rule. For instance, the Mamdani IF-THEN rule to infer the "$Overlaps^{(\alpha,\beta)}_{(1)}(I,J)$" relation is the following :

(define-concept Rule0 (g-and (some $Precede_{(1/1)}$ Fulfilled) (some $Precede_{(1/2)}$ Fulfilled) Fulfilled) (some $Precede_{(1/3)}$ Fulfilled) (some $Overlaps_{(1)}$ True))) // Fuzzy rule

We define three input fuzzy variables, named "Precede(1/1)", "Precede(1/2)" and "Precede(1/3)", which have the same MF than that of "$Precede^{(\alpha,\beta)}_{(1)}$". We define one output variable "Overlaps(1)" which has the same membership than that of the fuzzy object property "FuzzyRelationIntervals". "Precede(1/1)", "Precede(1/2)" and "Precede(1/3)" are instantiated with, respectively, $(I^{-(N)}-J^{-(1)})$, $(J^{-(N)}-I^{+(1)})$ and $(I^{+(N)}-J^{+(1)})$.

4.4 Conclusion

In this chapter, we proposed two approaches to represent and reason on imprecise time intervals in OWL : a crisp-based approach and a fuzzy-based approach. The crisp-based approach is based only on

TABLE 4.3 : Fuzzy gradual personalized temporal interval relations upon imprecise time intervals.

Relation	Inverse	Relations between bounds	Definition
$Before^{(\alpha,\beta)}_{(K)}(I,J)$	$After^{(\alpha,\beta)}_{(K)}(I,J)$	$\forall I^{+(i)} \in I^+, \forall J^{-(j)} \in J^- : (I^{+(i)} < J^{-(j)})$	$Precede^{(\alpha,\beta)}_{(K)}(I^{+(N)}, J^{-(1)})$
$Meets^{(\alpha,\beta)}(I,J)$	$MetBy^{(\alpha,\beta)}(I,J)$	$\forall I^{+(i)} \in I^+, \forall J^{-(j)} \in J^- : (I^{+(i)} = J^{-(j)})$	$Min(Same^{(\alpha,\beta)}(I^{+(1)}, J^{-(1)}) \wedge Same^{(\alpha,\beta)}(I^{+(N)}, J^{-(N)}))$
$Overlaps^{(\alpha,\beta)}_{(K)}(I,J)$	$OverlappedBy^{(\alpha,\beta)}_{(K)}(I,J)$	$\forall I^{-(i)} \in I^-, \forall I^{+(i)} \in I^+, \forall J^{-(j)} \in J^-, \forall J^{+(j)} \in J^+ : (I^{-(i)} < J^{-(j)}) \wedge (J^{-(j)} < I^{+(i)}) \wedge (I^{+(i)} < J^{+(j)})$	$Min(Precede^{(\alpha,\beta)}_{(K)}(I^{-(N)}, J^{-(1)}) \wedge Precede_{(K)}{}^{(\alpha,\beta)}(J^{-(N)}, I^{+(1)}) \wedge Precede^{(\alpha,\beta)}_{(K)}(I^{+(N)}, J^{+(1)}))$
$Starts^{(\alpha,\beta)}_{(K)}(I,J)$	$StartedBy^{(\alpha,\beta)}_{(K)}(I,J)$	$\forall I^{-(i)} \in I^-, \forall I^{+(i)} \in I^+, \forall J^{-(j)} \in J^-, \forall J^{+(j)} \in J^+ : (I^{-(i)} = J^{-(j)}) \wedge (I^{+(i)} < J^{+(j)})$	$Min(Same^{(\alpha,\beta)}(I^{-(1)}, J^{-(1)}) \wedge Same^{(\alpha,\beta)}(I^{-(N)}, J^{-(N)}) \wedge Precede^{(\alpha,\beta)}_{(K)}(I^{+(N)}, J^{+(1)}))$
$During^{(\alpha,\beta)}_{(K)}(I,J)$	$Contains^{(\alpha,\beta)}_{(K)}(I,J)$	$\forall I^{-(i)} \in I^-, \forall I^{+(i)} \in I^+, \forall J^{-(j)} \in J^-, \forall J^{+(i)} \in J^+ : (J^{-(j)} < I^{-(i)}) \wedge (I^{+(i)} < J^{+(i)})$	$Min(Precede^{(\alpha,\beta)}_{(K)}(J^{-(N)}, I^{-(1)}) \wedge Precede^{(\alpha,\beta)}_{(K)}(I^{+(N)}, J^{+(1)}))$
$Ends^{(\alpha,\beta)}_{(K)}(I,J)$	$EndedBy^{(\alpha,\beta)}_{(K)}(I,J)$	$\forall I^{-(i)} \in I^-, \forall I^{+(i)} \in I^+, \forall J^{-(j)} \in J^-, \forall J^{+(j)} \in J^+ : (I^{-(i)} < J^{-(j)}) \wedge (I^{+(i)} = J^{+(j)})$	$Min(Precede^{(\alpha,\beta)}_{(K)}(J^{-(N)}, I^{-(1)}) \wedge Same^{(\alpha,\beta)}(I^{+(1)}, J^{+(1)}) \wedge Same^{(\alpha,\beta)}(I^{+(N)}, J^{+(N)}))$
$Equal^{(\alpha,\beta)}(I,J)$	$Equal^{(\alpha,\beta)}(I,J)$	$\forall I^{-(i)} \in I^-, \forall I^{+(i)} \in I^+, \forall J^{-(j)} \in J^-, \forall J^{+(j)} \in J^+ : (I^{-(i)} = J^{-(j)}) \wedge (I^{+(i)} = J^{+(j)})$	$Min(Same^{(\alpha,\beta)}(I^{-(1)}, J^{-(1)}) \wedge Same^{(\alpha,\beta)}(I^{-(N)}, J^{-(N)}) \wedge Same^{(\alpha,\beta)}(I^{+(1)}, J^{+(1)}) \wedge Same^{(\alpha,\beta)}(I^{+(N)}, J^{+(N)}))$

crisp environment. We extended the 4D-fluents model to represent imprecise time intervals and crisp interval relations in OWL 2. To reason on imprecise time intervals, we extended the Allen's interval algebra in a crisp way. Inferences are done via a set of SWRL rules. The fuzzy-based approach is entirely based only on fuzzy environment. We extended the 4D-fluents model to represent imprecise time intervals and fuzzy interval relations in Fuzzy-OWL 2. To reason on imprecise time intervals, we extend the Allen's interval algebra in a fuzzy gradual personalized way. We infer the resulting fuzzy interval relations in Fuzzy-OWL 2 using a set of Mamdani IF-THEN rules.

Concerning the choice between these two approaches (the crisp-based one or the fuzzy-based one), as a fuzzy ontology is an extension of crisp ontology, researchers may choose any of our two approaches for introducing imprecise interval management in their knowledge bases whatever these latter are crisp or fuzzy. However our fuzzy-based approach provides more functionalities, in particular it is suitable to represent and reason on gradual interval relations such as "middle before" or "approximately at the same time". Hence, in the case of manipulating a fuzzy knowledge base, we encourage researchers to choose the fuzzy-based approach to model and reason on imprecise time intervals. The main interest of the crisp-based approach is that this solution can be implemented with classical crisp tools and that the programmers are not obliged to learn technologies related to fuzzy ontology. Considering that crisp tools and models are more mature and better support scaling, the crisp-based approach is more suitable for marketed products.

Our work have been studied in two projects, having in common to manage life logging data :

— The VIVA [2] project that aims to design the Captain Memo memory prosthesis [MGH+15, HHM+15] for Alzheimer Disease patients. Among other functionalities, this prosthesis manages a knowledge base on the patient's family tree, using an OWL ontology. Imprecise inputs are especially numerous when given by an Alzheimer Disease patient. Furthermore, dates are often given in reference to other dates or events. Thus we used, to represent this data, our "fuzzy" solution. One interesting point in this solution is to deal with a personalized slicing of the person's life in order to sort the different events.

2. http://viva.cnam.fr/

— The QUALHIS [3] project that aims to allow Middle Ages specialized historians to deal with prosopographical data bases storing Middle age academic's career histories. Data come from various archives among Europe and data about a same person are very difficult to align. Representing and ordering imprecise time interval is required to redraw the careers across the different European universities who hosted the person. We preferred the "crisp" solution in order to favour the integration within the existing crisp ontologies and tools, and to ease the management by Historians.

3. http://www.univ-paris1.fr/fileadmin/Axe_de_recherche_PIREH/aap2017-mastodons-Lamasse.pdf

Chapitre 5

Quality Assessment of KGs : Completeness and Conciseness

This chapter is based on work realized in the context of the PhD theses of Subhi Issa and Pierre-Henri Paris that I co-supervised with Dr. Samira Cherfi (from Cnam Paris). The result of this work was published in : [IPHC19a], [IPH17], [IHC19]

A high quality of data can help applications to produce more accurate results. Several approaches and tools have been proposed to evaluate varying dimensions of Knowledge Graphs quality. However, for the majority of dimensions, these approaches do not provide an in-depth study but only generic measures to evaluate data quality [ZRM+16]. In the context of Knowledge Graphs enrichment, one of the objective of interlinking data (or simply enriching data) is to be able to reuse information, i.e., to increase the completeness of an entity using one or more other entities. This quality dimension is recognized as important, since providing completeness information about a data source helps increasing the confidence of such a source. The completeness dimension can take two forms : *(i)* at the schema level of the entity, i.e., the number of different properties it uses, and *(ii)* at the data level, i.e., the number of values a property can have. Measuring data completeness is very difficult since it is almost impossible to establish a gold standard for data in the Semantic Web domain due to the open-world assumption ([DNPR13]). Thus, we propose in this work an approach that assess the completeness at the schema level. The first objective of our approach proposed is to generate a conceptual schema. This schema will allow measuring the completeness of the schema of an entity of the Knowledge Graph.

Another importante dimension that we address in this work is the conciseness. It is one of these dimensions that should be extensively examined as it prevents data redundancy. Indeed, wasteful data redundancy generally occurs when a given piece of data should not be repeated. It is important to identify this redundancy because, on the one hand, having unuseful data may influence the accuracy of information negatively, and on the other hand, this requires query expansion to extract all needed information. In this chapter, we present an approach that identifies repeated predicates to enhance the dataset conciseness. The objective is to find the equivalent predicates to remove relations which do not present new information to the Knowledge Graph. The proposed approach consists of three phases. It is based, in addition to a statistical analysis, on a semantic analysis through studying the meaning of each predicate to detect logical conflicts. We especially aim to analyze the meaning of the predicates in order to find those which are equivalent predicates, or exclude them when their semantic features are different. We take into account several semantic features such as symmetric, transitive, functional, inverse functional, disjoint domain/range and cardinality restrictions. Finally, we take into account the meaning of the predicate in the context using learning algorithms.

The remainder of this chapter is organized as follows : Section 5.1 presents related literature on KGs completeness assessment, details our mining-based approach for RDF data conceptual modeling, and presents two use cases for the LOD-CM, a web tool that we developed to validate the approach. Section 5.2 present related work about the conciseness dimension, describes a motivating example, explains our proposed approach that consists of three sequential phases, and finally presents two experiments performed on real-world datasets to evaluate our proposed approach that aims to assess conciseness of Knowledge Graphs. Finally, Section 5.3 draws conclusions.

5.1 Assessing Completeness of RDF Datasets

Data became a strategic asset in the information-driven world. One of the challenges for companies and researchers is to improve the display and understandability of the data they manage and use. However, exploiting and using Knowledge Graphs, even if it is more and more accessible, is not an easy task. Data is often incomplete and lacks metadata. These issues mean that the quality of published data is not as good as we could expect, leading to a low added value and low reliability of the derived conclusions. In

[JHY+10], the authors believe that existing approaches that describe Knowledge Graphs focus on their statistical aspects rather than on capturing conceptual information.

A conceptual schema is an abstraction of a reality that can serve as a vehicle for understanding, communicating, reasoning, and adding knowledge about this reality. In traditional information system development, conceptual modeling is driven by intended usage and needs. For Knowledge Graphs, as in all user-generated content, data is rather use-agnostic [LP15]. As a result, the data is represented according to many individual points of view. These points of view lead to a lack of semantics, whereas semantics is necessary for reasoning about the data. We believe that a conceptual schema that creates an abstract representation upon data would help to overcome the disparity of visions and will reveal the underlying semantics [Oli07]. Let us consider, for instance, that we have a collaboratively built knowledge graph. In this case, the traditional top-down vision of a predefined schema is no more applicable. Both data and underlying schema evolve continuously, as several communities describe data with different views and needs. In this situation, a conceptual schema, defined as an abstract and consensual representation about the reality that is derived from requirements, could not be applied. The challenge is then to find a way to create a suitable conceptual schema having entities as a starting point.

The **research questions** of this chapter are as follows :

— How to compute the schema completeness of an entity in an RDF-based knowledge graph ?

— How to facilitate access to the information structure of a knowledge graph ?

In this chapter, we are interested in the conceptual modeling of RDF-based Knowledge Graphs [KC04]. Our objective is to define an approach for deriving conceptual schemas from existing data. The proposed solution should cope with the essential characteristics of a conceptual schema that are the ability to make an abstraction of relevant aspects from the universe of discourse and the one of meeting user's requirements [RP00]. The approach we propose in this chapter takes into account the two facets, namely the universe of discourse represented by the data sources, and the user's needs represented by the user's decisions during the conceptual schema construction. As the model should express the meaningful state of the considered knowledge graph, we rely on a mining approach leading to taking into consideration the data model from a more frequent combination of properties. The relevancy of properties is handled by integrating a completeness measurement solution that drives the identification of relevant properties. To meet user's requirements, we propose to construct the conceptual schema by allowing the user to decide about the parts of the conceptual schema to reveal according to her needs and constraints.

The main contributions are :

(1) We use a mining approach to infer a model from data, as we consider that no predefined schema exists. The underlying assumption is that the more frequent a schema is, the more representative for the knowledge graph it is.

(2) We introduce a novel approach, called *LOD-CM*, for conceptual schema mining based on quality measures, and, in this chapter, on completeness measures as a way to drive the conceptual schema mining process.

5.1.1 Related work

RDF data is described as sets of statements called *triples*. A triple $\langle s, p, o \rangle$ is a fact where a subject s has a property p, and the property value is the object o. As an example, $\langle England, capital, London \rangle$ means that London is the capital city of England. Understanding and reasoning about this data require at least knowledge about its abstract model. Consequently, schema discovery has attracted several researchers originating from several communities. The research directions address objectives such as efficient storage, efficient querying, navigation through data or semantic representation, etc.

Completeness of Knowledge Graphs is one of the most critical data quality dimensions ([BS16]). This dimension is defined as the degree to which all required information is present in a particular knowledge graph ([ZRM+13]). We have to know that a reference schema (or a gold standard) should be available to compare against a given knowledge graph.

In the database community, the question was how to store this kind of data. [LM09] proposed a solution

that derives a classical relational schema from an RDF data source to accelerate the processing of queries. In the FlexTable method ([WDLW10]), authors proposed to replace RDF triples by RDF tuples resulting from the unification of a set of triples having the same subject. All these approaches do not target a human-readable schema and are more concerned with providing a suitable structure for a computer processing of data.

The Semantic Web community is more aware of data semantics through the usage of *ontologies* and *vocabularies*. Several semi-automatic or automatic proposals, mainly based on classification, clustering, and association analysis techniques are proposed. In [VN11] a statistical approach based on association rules mining allows generating ontologies from RDF data. Other works, such as those presented in [CPF15, PPEB15, KK15], are closer to modeling. The authors propose to derive a data structure using a clustering algorithm. After manual labeling of clusters representing groups of frequent properties, a schema is derived. These approaches, however, do not consider the user's needs and preferences, and the derived schema is the result of automatic preprocessing, apart from the labeling task.

In traditional conceptual modeling, models are generally derived from user's requirements. However, with the increasing use of external data sources in information systems, there is a need to apply bottom-up modeling from entities. This is motivated by the expressiveness and the analysis facilities that conceptual schemas could provide for such data. Similarly to our bottom-up approach, [LPS19] proposed a conceptual modeling grammar based on the assumption that entities play a major role while human beings try to represent reality. In [LP15], the authors presented a set of principles for conceptual modeling within structured user-generated content. The authors highlighted the problem of quality in such produced content. They focused on the importance of capturing relevant properties from entities. However, the proposal does not provide an explicit solution for deriving such models. Concerning unstructured data, we can cite [EL13], where authors addressed the problem of deriving conceptual schemas based on regular-expression pattern recognition.

Recognizing that conceptual modeling is a powerful tool for data understanding, our proposal addresses the problem of deriving a conceptual schema from RDF data. By exploring entities, our approach integrates a completeness measurement as a quality criterion to ensure the relevancy of the derived schema as data from RDF data sources is the result of a free individual publication effort. The result would be a conceptual schema enriched with completeness values.

5.1.2 Conceptual schemas derivation

To illustrate our proposed approach, let us consider a user willing to obtain a list of artists with their names and birthplaces from an RDF-based knowledge graph ; To do so, she can write the following SPARQL query [1] :

```
SELECT * WHERE
{
  ?actor rdf:type dbo:Actor .
  ?actor foaf:name ?name .
  ?actor dbo:birthPlace ?birthPlace .
}
```

Listing 5.1 : SPARQL query retrieving all actor names and birthplaces.

Writing such a query is much more difficult in a Linked Open Data (LOD) source context than in a relational database one. In a relational context, the database schema is predefined, and the user writing the query is aware of it. With Knowledge Graphs, the schema does not exist. Moreover, there is another problem related to data completeness : The expressed query returns only the list of actors having values for all the properties listed in the query. In our example, only actors having values for both *foaf:name* and

1. Performed on : `http://dbpedia.org/sparql`

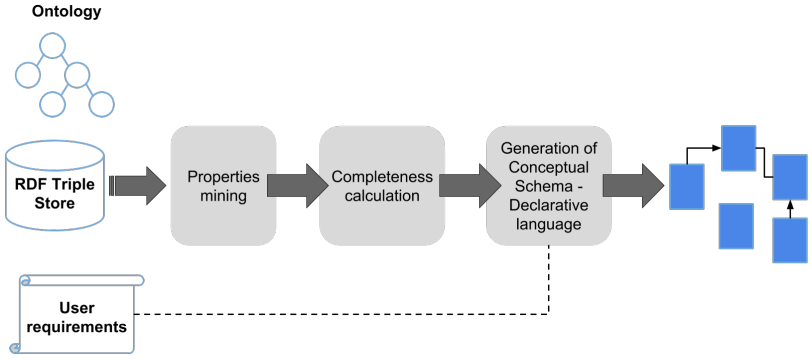

FIGURE 5.1 : The *LOD-CM* Workflow

dbo:birthPlace are included in the result. Knowing that at most 74% of actors have a value for *dbo:birthPlace*, the user should probably appreciate getting this information to add, for example, `OPTIONAL` to the second pattern of the query and obtain more results. Besides, she would be aware of the fact that the result is complete to a certain degree (i.e., *dbo:birthPlace* is present in only 74% of actors).

To tackle these two problems, we propose an approach that aims to help "revealing" a conceptual schema from an RDF-based knowledge graph. This conceptual schema is driven by the user for both its content and completeness quality values.

In the context of the Web of Data, most of the Knowledge Graphs published in the Web are described by models called, in Knowledge Graphs jargon, vocabularies (or ontologies). However, these models are not used in a prescriptive manner. Consequently, a person who publishes data is not constrained by the underlying ontology leading to sparse descriptions of concepts. For example, the instances of class *Actor* from DBpedia use around 532 properties that are not equally relevant.

From these observations, it is clear that checking data (entities) is necessary to infer a relevant model that can be used to guarantee, for example, an expected completeness value. The approach that we propose deals with this issue through an iterative process, which infers a conceptual schema complying with the expected completeness. Figure 5.1 gives an overview of this process.

The process of inferring a conceptual schema goes through four steps : First, a subset of data that corresponds to the user's scope is extracted from the knowledge graph (cf. Section 5.1.2). This subset is then transformed into transactions, and a mining algorithm is applied. In our approach, for efficiency reasons, we chose the well-known FP-growth algorithm [HPY00, HPYM04] (any other itemset mining algorithm could obviously be used). From the generated frequent itemsets, only a subset of these frequent itemsets, called "Maximal" [Jr.98, GZ01, GZ03], is captured. This choice is motivated by the fact that, on the one hand, we are interested in the *expression* of the frequent pattern, and, on the other hand, the number of frequent patterns could be exponential when the transaction vector is huge (cf. Section 5.1.2). $\mathcal{MFP}$ is the set containing all maximal frequent patterns. Next, each pattern in $\mathcal{MFP}$ is used to calculate the completeness of each transaction. The presence or absence of the pattern is reflected in the completeness. Hence, the completeness of the whole knowledge graph regarding this pattern can be computed by aggregating transaction completenesses. The final completeness value will be the average of all completeness values calculated for each $\mathcal{MFP}$ pattern (cf. Section 5.1.2). Finally, based on the completeness value and $\mathcal{MFP}$ that guarantees this value, a conceptual schema is generated. The classes, the attributes, and the relations of the model will be tagged with the completeness value (cf. Section 8). All these steps are integrated into an iterative process : the user could choose some parts in the generated model to refine. The data corresponding to the parts to refine is then extracted from the knowledge graph, and the same steps are carried out to generate a new model.

In the following subsections, we give a detailed description of each step of the workflow.

Scope and Completeness Specification

In this step, a subset of data is extracted from the knowledge graph. This subset could correspond to a class or a set of classes such as *Actor*, *Film* , or *Organization*. The subset defines what we call the user's scope that corresponds to the classes that the user plans to use in a query, or to the information she wants to explore or any kind of usage based on data consumption.
The user is also asked to indicate the degree of the desired completeness. Indeed, properties for a given class are not equally used. For example, for the class *Artist*, the property *foaf:name* has a value for 99% of the entities, whereas the *dbo:birthPlace* property has a value for at most 74% of the entities. Our approach gives the possibility to express a constraint on the completeness values desired for mined properties and associations. Once the classes are identified, the data is converted into transaction vectors, and a mining algorithm is applied to obtain a set of frequent itemsets.
Table 5.1 illustrates some entities of the *Film* class in the form of triples, taken from DBpedia. Each class is described by a set of properties (predicates). An entity of this class could have a value for all or a subset of these properties. This subset is called a transaction. Table 5.2 represents the set of transactions constructed from the triples of Table 5.1.
More formally, given a knowledge graph $\mathcal{KG}$, let us define a set of classes $C \in \mathcal{KG}$ (e.g., *Film*, *Artist*), $\mathcal{E}_C \in \mathcal{KG}$ is the set of entities for classes in C (e.g., *The_Godfather* is an entity of the *Film* class), and $P_C = \{p_1, p_2, ..., p_n\} \in \mathcal{KG}$ is the set of properties used by entities in $\mathcal{E}_C$ (e.g $director(Film, Person)$).
Given a subset of entities $E = \{e_1, e_2, ..., e_m\}$ with $E \subseteq \mathcal{E}_C$ (e.g., properties used to describe the *The_Godfather* entity are : *director* and *musicComposer*), $\mathcal{T}_E = (t_1, t_2, ..., t_m)$ is the list of transactions where $\forall k, 1 \leq k \leq m$: $t_k \subseteq P_C$ and t_k is the set of properties used in the description of $e_k \in E$, i.e., $\forall p \in t_k, \exists o \in \mathcal{KG} : \langle e_k, p, o \rangle \in \mathcal{KG}$.
We consider $\mathcal{CP}$ the completeness of E against properties used in the description of each of its entities. Moreover, $\mathcal{P}(t_k)$ is the power set of transaction t_k.

Properties Mining

All statements having a subject from a class C_1 are grouped. The related properties of those statements could consequently constitute either the attributes (properties) of the class C_1 or relationships to other classes when the property value (the object in the triple $< s, p, o >$) refers to another class. In this step, the objective is to find the properties patterns that are the most shared by the subset of entities extracted from the knowledge graph. This set will be then used to calculate a completeness value regarding these patterns.
Let $C, \mathcal{E}_C, P_C$ be respectively classes, instances (of classes in C), and properties (of entities in $\mathcal{E}_C$) of a knowledge graph $\mathcal{KG}$ and E be a subset of data (entities) extracted from $\mathcal{KG}$ with $E \subseteq \mathcal{E}_C$. We first initialize $\mathcal{T}_E = \emptyset$, $\mathcal{MFP} = \emptyset$. For each $e \in E$, we generate a transaction t, i.e., properties used by e. Indeed, each entity e is related to values (either resources or literals) through a set of properties. Therefore, a transaction t_k of an entity e_k is a set of properties such that $t_k \subseteq P_C$. Transactions generated for all the entities of E are then added to the $\mathcal{T}_E$ list.
Example 5.1.1 : Referring Table 5.1, let E be a subset of entities such that :
$E = \{The\_Godfather, Goodfellas, True\_Lies\}$. The list of transactions $\mathcal{T}_E$ would be :

$$\mathcal{T}_E = (\{director, musicComposer\}, \{director, editing\}, \{director, editing, musicComposer\})$$

The objective is then to compute the set of frequent patterns $\mathcal{FP}$ from the transaction vector $\mathcal{T}_E$.
Definition 5.1.1 : (**Pattern**) Let $\mathcal{T}_E$ be a set of transactions. A pattern $\hat{P}$ is a sequence of properties shared by one or several transactions t in $\mathcal{T}_E$. It is sometimes called an *itemset*.

For any pattern $\hat{P}$, let $\mathcal{P}(\hat{P})$ be the power set of $\hat{P}$ (composed, in our case, of properties), and $T(\hat{P}) = \{t \in \mathcal{T}_E \mid \mathcal{P}(\hat{P}) \subseteq \mathcal{P}(t)\}$ be the corresponding set of transactions. $\mathcal{P}(\hat{P})$ designates the *expression* of $\hat{P}$, and $|T(\hat{P})|$

Subject	Predicate	Object
The_Godfather	director	Francis_Ford_Coppola
The_Godfather	musicComposer	Nino_Rota
Goodfellas	director	Martin_Scorsese
Goodfellas	editing	Thelma_Schoonmaker
True_Lies	director	James_Cameron
True_Lies	editing	Conrad_Buff_IV
True_Lies	musicComposer	Brad_Fiedel

TABLE 5.1 : A sample of triples from DBpedia

entity	Transaction
The_Godfather	director, musicComposer
Goodfellas	director, editing
True_Lies	director, editing, musicComposer

TABLE 5.2 : Transactions extracted from triples

the *support* of $\hat{P}$. A pattern $\hat{P}$ is frequent if $\frac{1}{|\mathcal{T}_E|}|T(\hat{P})| \geq \xi$, where ξ is a user-specified threshold.

Example 5.1.2 : Referring Table 5.2, let $\hat{P} = \{director, musicComposer\}$ and $\xi = 60\%$. $\hat{P}$ is frequent as its relative support (66.7%) is greater than ξ.

To find all the frequent patterns $\mathcal{FP}$, we used, as we mentioned above, the FP-growth itemsets mining algorithm. However, according to the size of the transactions vector, the FP-growth algorithm could generate a very large $\mathcal{FP}$ set. As a reminder, our objective is to see how a transaction (a description of an entity) is *complete* against a set of properties. Thus, we focus on the pattern *expression* (in terms of items it contains) instead of its *support*.

For completeness calculation, we need to select a pattern to serve as a reference schema. This pattern should present the right balance between frequency and expressiveness. Therefore we use the concept, called "Maximal" frequent patterns, to find this subset. Thus, to reduce $\mathcal{FP}$, we generate a subset containing only "Maximal" patterns.

Definition 5.1.2 : ($\mathcal{MFP}$) Let $\hat{P}$ be a frequent pattern. $\hat{P}$ is maximal if none of its proper superset is frequent. We define the set of Maximal Frequent Patterns $\mathcal{MFP}$ as :

$$\mathcal{MFP} = \{\hat{P} \in \mathcal{FP} \mid \nexists \hat{P}' \in \mathcal{FP} : \hat{P} \subset \hat{P}' \wedge \frac{|T(\hat{P}')|}{|\mathcal{T}_E|} < \xi\}$$

Example 5.1.3 : Referring Table 5.2, let $\xi = 60\%$ and the set of frequent patterns $\mathcal{FP} = \{\{director\}, \{musicComposer\}, \{editing\}, \{director, musicComposer\}, \{director, editing\}\}$. The $\mathcal{MFP}$ set would be :

$$\mathcal{MFP} = \{\{director, musicComposer\}, \{director, editing\}\}$$

Completeness calculation

In this step, we carry out for each transaction a comparison between its corresponding properties and each pattern of the $\mathcal{MFP}$ set (regarding the presence or the absence of the pattern). An average is, therefore, calculated to obtain the completeness of each transaction $t \in \mathcal{T}_E$. Finally, the completeness of the whole $t \in \mathcal{T}_E$ will be the average of all the completeness values calculated for each transaction.

Definition 5.1.3 : (*Completeness*) Let E be a subset of entities, $\mathcal{T}_E$ the set of transactions constructed from E, and $\mathcal{MFP}$ a set of maximal frequent pattern. The completeness of E corresponds to the completeness of its transaction vector $\mathcal{T}_E$ obtained by calculating the average of the completeness of $\mathcal{T}_E$ regarding each

pattern in $\mathcal{MFP}$. Therefore, we define the completeness $\mathcal{CP}$ of a subset of entities E as follows :

$$\mathcal{CP}(E) = \frac{1}{|\mathcal{T}_E|} \sum_{k=1}^{|\mathcal{T}_E|} \sum_{j=1}^{|\mathcal{MFP}|} \frac{\delta(\hat{P}_j, \mathcal{P}(t_k))}{|\mathcal{MFP}|} \tag{5.1}$$

such that : $\hat{P}_j \in \mathcal{MFP}$, and

$$\delta(\hat{P}_j, \mathcal{P}(t_k)) = \begin{cases} 1 & \text{if } \hat{P}_j \subset \mathcal{P}(t_k) \\ 0 & \text{otherwise} \end{cases}$$

Algorithm 4 shows the pseudo-codes for calculating $\mathcal{CP}(E)$.

Example 5.1.4 : Let ξ = 60%. The completeness of the subset of entities in Table 5.1 regarding $\mathcal{MFP}$ = {{*director,musicComposer*}, {*director,editing*}} *would be :*

$$\mathcal{CP}(E) = \frac{2 \times (1/2) + (2/2)}{3} = 0.67$$

This value corresponds to the completeness average value for the whole knowledge graph regarding the inferred patterns in $\mathcal{MFP}$.

input : $\mathcal{KG}$, E, ξ
output : $\mathcal{CP}(E)$

1 **foreach** $e \in E$ **do**
2 $\quad t_i = \left|p_1 \quad p_2 \quad \dots \quad p_n\right|$;
3 $\quad \mathcal{T}_E = \mathcal{T}_E + t_i$;
4 **end**
5 /* *Properties mining* */
6 $\mathcal{MFP} = Maximal(FP\text{-}growth(\mathcal{T}_E, \xi))$;
7 /* *Using equation 5.1* */
8 **return** $CP(E) = CalculateCompleteness(E, \mathcal{T}_E, \mathcal{MFP})$

Algorithm 4 : Completeness calculation

Generation of Enriched Conceptual Schemas

In this step, the goal is to generate a conceptual schema enriched with the completeness values calculated in the previous step. The $\mathcal{MFP}$ patterns used to get the completeness values are transformed into a class diagram. Figure 5.2 illustrates the user's interface of our LOD-CM web service. Using the graphical interface [2], the user can choose her constraints. The web service permits the user to enter the class name in the text box, and the user may select the threshold completeness she wants to apply. Currently, our demo supports DBpedia and Wikidata Knowledge Graphs.

After the user selects the class name and desired completeness and clicks the "Submit" button, the algorithm runs to find the attributes, relationships, and the missed domains/ranges based on user's constraints.

The structure of the model is constructed regarding the definitions of the properties of patterns in the ontology. Figure 5.3 represents a class diagram derived by our approach, from a set of films extracted from DBpedia.

In this example, the expectation of the user is a model that guarantees at least 50% of completeness. To generate the model, the first step consists of obtaining the set of properties $p \in \bigcup_{j=1}^{n} \mathcal{P}(\hat{P}_j)$, and $\hat{P}_j \in \mathcal{MFP}$ that composes the union of all the $\mathcal{MFP}$, mined from the extracted subset, with a minimum support ξ = 50%. For this example, the set of properties are : {*director, label, name, runtime, starring, type*}, {*director, label,*

2. http://cedric.cnam.fr/lod-cm

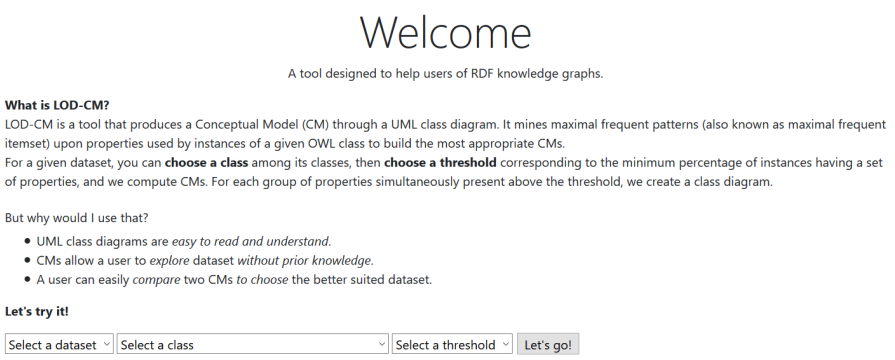

FIGURE 5.2 : LOD-CM main interface

name, starring, type, writer} and {*label, name, runtime, type, writer*}. OWL 2 distinguishes between two main classes of properties : *(i)* datatype properties, where the value is a literal, and *(ii)* object properties, where the value is an individual (i.e., another entity of a different class). Each property is considered as an attribute (e.g., name) of the class or a relationship (e.g., director) with another class. Two types of links will be used during the generating of conceptual schemas : inheritance and association links. Inheritance link describes the relationship between the class and the superclass, and the association link describes the relationship between two classes and points to the property. A dotted link was added to illustrate that a class has been inferred to complete the relationship. For this reason, based on the approach that has been proposed in [TKS12a], we infer missed domains (and ranges) of properties. In our example, the class names and the inheritance links between the classes are derived from class names and subsumptions described in the DBpedia ontology. We do not derive new class names nor new subsumptions as the conceptual schema should conform to the data used. Indeed, even if the derived conceptual schema is not satisfactory from conceptual modeling principles, it should faithfully reflect the reality of data while taking into account user preferences. Finally, the diagram is enriched by the completeness values calculated in the previous step. These values are associated with each component of the model.

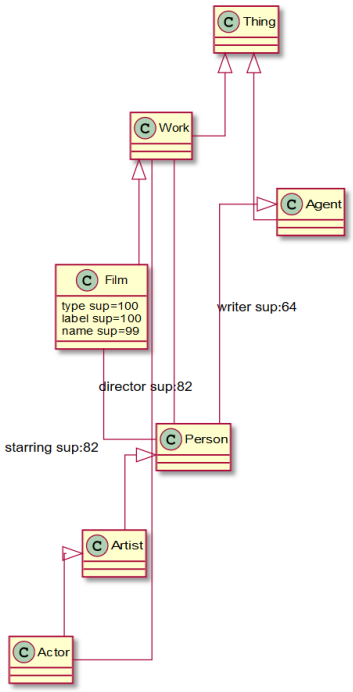

FIGURE 5.3 : The *Film* conceptual schema as a class diagram

A new iteration is triggered when the user chooses to get more details about a part of the model (e.g., the

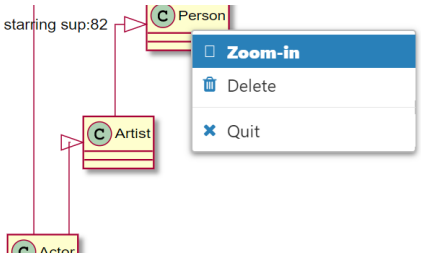

FIGURE 5.4 : Contextual menu for navigation and editing.

Artist class, see Fig. 5.4). In this case, a new query is executed on the knowledge graph to extract data corresponding to this part. The previous three steps are then executed to generate a new model integrating the new desired details. Figure 5.5 shows an example that details a part of the model from Figure 5.3. In this example, a set of classes, relationships, and attributes are added to the *Artist* class with corresponding completeness values. This way of revealing the conceptual schema is similar to a magnifying glass that allows the user to navigate around a targeted concept, here the *Film* class.

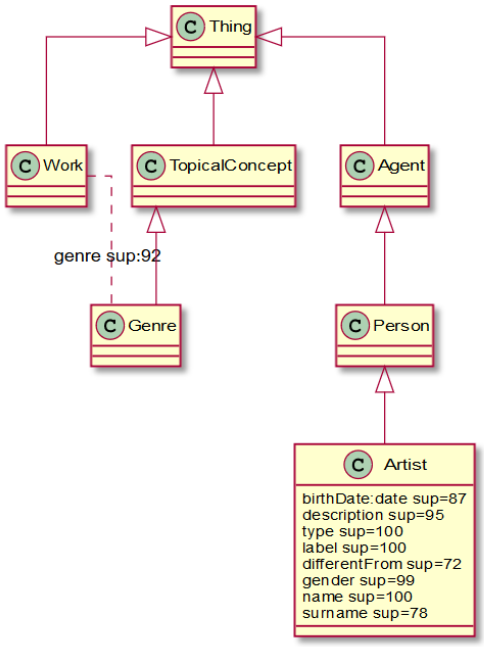

FIGURE 5.5 : The *Artist* diagram class

The output of our algorithm is a file written in a declarative language. The file includes the chosen class, the attributes, and the relationships tagged by completeness values. We use PlantUML [3] to transfer this generated file into a picture to illustrate it to the user.

5.1.3 Use cases

The objective of the Linked Open Data cloud is to enable large-scale data integration so that we can have a contextually relevant Web and find quick answers to a much wider range of questions. LOD-CM is a web-based completeness demonstrator for RDF-based Knowledge Graphs. It is used to display data related to the chosen class of a knowledge graph. In this section, we provide a summary of two use cases related to schema discovery based on user's needs. The displayed model could help the user to understand the schema and discover the related properties. LOD-CM only supports two Knowledge Graphs that are DBpedia and Wikidata.

3. `http://plantuml.com/`

Class/threshold	0.1	0.3	0.5	0.7	0.9
Film	18	12	7	6	3
Settlement	18	14	8	5	4
Organisation	18	4	4	3	3
Scientist	19	16	12	9	5

TABLE 5.3 : DBpedia number of predicates by classes and thresholds

Class diagram to facilitate data browsing

LOD-CM aims to visualize the discovered schema based on the user's requirements. Suppose a user wants to find the directors and budgets of a list of films. 82% of films have a director in the DBpedia knowledge graph. Besides, only 15% of films have budget value for the same knowledge graph. Only the list of films that have the properties (director and budget) will be displayed (i.e., at most 15% of the films). The outcome model could help the user to present the properties that are related to the chosen class in a proportion greater than a specified threshold. Besides, it illustrates the relationship between the concerned classes. For example, the classes *Person* and *Film* are linked by the property *director*. Furthermore, the model illustrates the inheritance relationship, such as *Artist* is a subclass of *Person*.

Discovering a subset of MFP

As mentioned in Section 5.1.2, our goal is also to find the set of properties that can be used together in the query and does not exceed the selected threshold. For example, for the *Film* class with 60% of completeness, four sets of properties are greater than 60% {{type, name, label, director, writer}, {type, name, label, director, runtime}, {type, name, label, director, starring},{type, name, label, runtime, starring}}. For this reason, our LOD-CM interface enables the user to check the desired properties that appear in the returned model. It should be noted that the property which does not achieve the completeness threshold with other selected properties will be inactivated, such as *starring* and *writer* in our previous example. This case could help the user to be sure that the returned results for its query with this set of properties are equal or greater than the desired threshold.

Finally, Table 5.3 shows the number of properties we get (at the end of our pipeline) for several classes according to several thresholds. The lower the threshold is, the more properties there are. Thus, lower thresholds produce more complex conceptual schemas but with more noise. Hence, this tool can help the user to find the right balance between those two. More detailed experiments of our completeness evaluation approach are presented in the following paper : [IPH17]

5.2 Assessing the Conciseness of Knowledge Graphs

In this section, we address the problem of identifying synonym predicates for Knowledge Graphs. We propose an approach to discover equivalent properties to evaluate the conciseness dimension of Knowledge Graphs.

5.2.1 Related work

Ontology alignment is the process of identifying that several vocabularies are semantically related, which may be in one or more datasets. Different approaches [KS03, RB01] are applied to discover synonym objects. In addition to research in linking RDF at the conceptual and instance levels that have been investigated in the recent past, property alignment has not received much attention yet. There are some studies on mapping properties across RDF sources[FWJY12, GTJ+13, ZGB+17]. In [FWJY12], the authors identified similar relations using subject and object overlap of predicates. On the other hand, [GTJ+13] used also statistical measures to identify strictly equivalent relations rather than similarity in general. In [ZGB+17, GTHS15], authors provided methods to group equivalent predicates by clustering synonymous relations. PARIS [SAS11] is another well known approach which needs to be mentioned in this context. It combines instance and schema matching using probabilities with high accuracy. This work focuses

particularly on aligning two ontologies.

Our work is motivated by the problems we met when we tried to reveal conceptual schemas of RDF data soucres [IPHC19b]. We relied on a mining approach taking into consideration the data model from a more frequent combination of predicates based on completeness measurement [IPH17]. The idea is how to deal with predicates that have the same meaning with different names. Besides to completeness dimension, conciseness is one aspect of Knowledge Graphs quality dimensions [ZRM+16] which basically aims to avoid repetition elements. The eliminating of the synonymously used predicates aims to optimize the dataset to speed up processing. Mendes et al. [MMB12] categorized conciseness dimension into intensional and extensional conciseness. The first type, which is the intensional conciseness, measures a number of unique dataset attributes to the total number of schema attributes, thus this measurement is represented on the schema level. In a similar manner but on the instance level, extensional conciseness measures the number of unique objects to the real number of objects in the dataset. In [LUM07], the authors proposed an assessment metric by detecting the number of ambiguous instances according to those of semantic metadata sets in order to discover the duplicated instances. In the similar sense but under a different name, Füber et Hepp [FH11] defined the elements of representation like (classes, predicates and objects) under the domination of "uniqueness". Their definition suggested uniqueness of breadth at the schema level and uniqueness of depth at the instance level.

Indeed, the proposed metrics of conciseness assessment are based on a simple ratio, which compares the proportion of existing elements to the overall ones. Whereas, the conciseness at schema level is measured by the number of unique predicates and dataset classes to the total number of those elements in a schema [MMB12]. On instance level, it measures the number of unique instances to the total number of instances in the dataset [MMB12, FH11, LUM07]. In [AN13], the authors proposed an algorithm to discover the synonym predicates for query expansions. This algorithm is based on mining similar predicates according to their subjects and objects. However, their approach proposed a lot of predicates that are not synonyms (false positive). In this work, we want to focus on extracting equivalent predicates from RDF dataset.

5.2.2 Motivating scenario

Semantic data is usually collected from heterogeneous sources through a variety of tools for different purposes [LSL+06, Mik05]. Unfortunately, this sort of mixing could lead to decreasing the quality of data, so that we have proposed this approach to evaluate the conciseness dimension of Knowledge Graphs.

Our research on the conciseness dimension was inspired by Abedjan and Naumann's work "Synonym Analysis for Predicate Expansion" [AN13]. The authors proposed a data-driven synonym discovery algorithm for a predicate expansion by applying both schema analysis and range content filtering.

Range content filtering aims to represent a transaction as a distinct object with several predicates. For example, the object *Lyon*[4] city is connected with several predicates such as (*birthPlace*, *deathPlace* and *location*). The authors supposed that synonym predicates share a similar group of object values. For this reason, the proposed approach seeks the frequent patterns of predicates that share a significant number of object values.

In fact, it is not sufficient to synonymously discover the used predicates depending only on range content filtering. For example, the predicates *birthPlace* and *deathPlace* share significant co-occurrences with the same object values but they are definitely used differently. For this reason, the authors added another filter called "schema analysis" in order to overcome this problem. This filter is better in finding suitable synonym predicates. The authors supposed that the synonym predicates should not co-exist for the same instance. According to schema analysis, transactions of distinct subjects with several predicates are represented. By applying negative association rules [BMS97], the synonym predicates appear in different transactions. For example, the subject *Michael_Schumacher* does not have two synonym predicates such as *born* and *birthPlace* in the same dataset.

Now, we clarify the drawbacks of Abedjan and Naumann's approach through applying the following example (see Table 5.5). We use a sample of facts from DBpedia dataset to discover the synonym predicates.

4. Lyon is a French city

TABLE 5.4 : Six configurations of context and target [AN11].

Conf.	Context	Target
1	Subject	Predicate
2	Subject	Object
3	Predicate	Subject
4	Predicate	Object
5	Object	Subject
6	Object	Predicate

TABLE 5.5 : Facts in SPO structure from DBpedia.

Subject	Predicate	Object
Adam_Hadwin	*type*	*GolfPlayer*
Adam_Hadwin	*birthPlace*	*Moose_Jaw*
Adam_Hadwin	*nationality*	*Canada*
White_River	*sourceCountry*	*Canada*
White_River	*riverMouth*	*Lake_Superior*
White_River	*state*	*Ontario*

Based on range content filtering (Conf. 6 as illustrated in Table 5.4), all the predicates will be gathered into groups by each distinct object. Thus, in order to retrieve frequent candidates, results could be as in Table 5.6.

As a result, we can see that *nationality* and *sourceCountry* are already in the same transaction. By applying FP-growth algorithm [HPYM04], or any other itemset mining algorithm, for mining frequent itemsets, *nationality* and *sourceCountry* will be found as a frequent pattern.

The next step is to perform schema analysis (Conf. 1 as illustrated in Table 5.4) by considering subjects as a context to get transactions as illustrated in Table 5.7. By applying negative association rules, Abedjan and Naumann's algorithm shows that there is no co-occurrence between *sourceCountry* and *nationality* predicates. Therefore, it will propose *nationality* and *sourceCountry* as a synonym predicate pair, which is not correct because *nationality* and *sourceCountry* have different intentions.

5.2.3 Discovering synonym predicates

In the next subsections, we explain our proposed approach that consists of three phases. In addition to the statistical study through schema analysis and range content filtering, we basically intend to perform a semantic analysis to understand the meaning of the candidates. Finally, we use learning algorithms to filter the results of the two previous phases.

TABLE 5.6 : Range content filtering.

Object	Predicate
GolfPlayer	*type*
Moose_Jaw	*birthPlace*
Canada	*nationality*, *sourceCountry*
Lake_Superior	*riverMouth*
Ontario	*state*

TABLE 5.7 : Schema analysis.

Subject	Predicate
Adam_Hadwin	*type*, *birthPlace*, *nationality*
White_River	*sourceCountry*, *riverMouth*, *state*

Phase 1 : statistical analysis

As we have already mentioned, our goal is to start with statistical analysis in order to discover potential equivalent predicates. We are interested, in this part, in studying the appearance of each predicate by finding the frequent pattern with negative association rules. This part is basically inspired from Abedjan and Naumann's work [AN13] which proposed a data-driven synonym discovery algorithm for predicate expansion. Based on mining configuration of contexts and targets [AN11], the authors applied Conf. 1 and Conf. 6 as illustrated in Table 5.4 that represents schema analysis and range content filtering, respectively.

We extent the method that is explained in Section 5.2.2 to be suitable not only to generate candidate pairs of synonym predicates, but also to remove those that are actually not by semantic analysis. In the next subsection, we look forward to study the candidates depending on semantics features to decrease the number of predicates by identifying strictly equivalent predicates and eliminating non-equivalent ones.

Phase 2 : semantic analysis

Actually, some predicates are not easy to understand, share the same meaning with different identifiers, or have several meanings. For these reasons, calculating string similarity or synonym based measurements on predicate names alone does not suffice. Indeed, the first phase proposes candidate pairs as synonyms but also too many false positive results, especially in the case of large cross-domain datasets. As the previous example illustrated in Section 5.2.2, the predicates *nationality* and *sourceCountry* could have the same object (Conf. 6) such as *Canada*. They also never co-occur together for the same subject (Conf. 1). However, *nationality* is a predicate of an instance that its type is *Person* class and *sourceCountry* is a predicate of an instance that its type is *Stream* class. Thus, they should not be considered as synonyms as clarified below.

We add an extension to Abedjan and Naumann's work by studying the meaning of each candidate. Indeed, we examine the semantic representations of the synonym candidates that, under certain conditions, provide us with useful conclusions ; for example, a predicate could not be equivalent to another predicate if they have disjoint domains or ranges. Taking the previous example of *nationality* and *sourceCountry* predicates, according to the DBpedia ontology, *Stream* class is a subclass of *Place* class, and *Place* and *Person* classes are completely disjointed. As a consequence, we cannot consider *nationality* and *sourceCountry* to be equivalent predicates.

Thus, in this phase we take into account the semantic part of the predicates. This allows us to detect the incompatibility of the predicates that have opposite features such as symmetric and asymmetric. OWL2 supports declaring two classes to be disjointed. It also supports declaring that a predicate is symmetric, reflexive, transitive, functional, or inverse functional. We take into account these features for each predicate in addition to the *max* cardinality restriction.

We prove the disjointness of predicates, that could not be synonyms, using $\mathcal{SROIQ}$ description logic that models constructors which are available in OWL 2 DL [HKS06]. We depend on studying the meaning of predicates by analyzing their features.

Domain and range disjointness

In the following paragraphs, we give an example about the disjointness of domain and range between two predicates. For a given RDF triple, the property *rdfs:domain* indicates the class that appears as its subject and the property *rdfs:range* indicates the class or data value that appears as its object.

— **Domain of property**
We use here the property *rdfs:domain* to check whether the domains of the two compared predicates

are disjointed or not. If yes, we can state that these predicates cannot be synonyms.

Theorem 5.2.1 : Let p_1 & p_2 be two predicates and C_1 & C_2 be two classes, p_1 & p_2 cannot be synonyms if :

$$\exists p_1.\top \sqsubseteq C_1 \tag{5.2}$$
$$\exists p_2.\top \sqsubseteq C_2 \tag{5.3}$$
$$C_1 \sqcap C_2 \sqsubseteq \bot \tag{5.4}$$

Proof 5.2.1 : *Assume* $\exists x$, *that :*

$$p_1(x, y_1) \tag{5.5}$$
$$p_2(x, y_2) \tag{5.6}$$

We assert that :

$$(5.2) + (5.5) \Rightarrow C_1(x) \tag{5.7}$$
$$(5.3) + (5.6) \Rightarrow C_2(x) \tag{5.8}$$
$$(5.7) + (5.8) \Rightarrow C_1 \sqcap C_2 \not\sqsubseteq \bot \tag{5.9}$$
$$(5.4) + (5.9) \Rightarrow \bot \; absurd \tag{5.10}$$

As a result, we conclude that predicates that have disjointed domains are disjointed. In the same manner, we prove the other features discussed previously.

— **Range of property**

For a given RDF triple, the property *rdfs:range* indicates the class or the data value that appears as its object (the predicate range).

Theorem 5.2.2 : Let p_1 & p_2 be two predicates and C_1 & C_2 be two classes, p_1 & p_2 cannot be synonyms if :

$$\top \sqsubseteq \forall p_1.C_1 \tag{5.11}$$
$$\top \sqsubseteq \forall p_2.C_2 \tag{5.12}$$
$$C_1 \sqcap C_2 \sqsubseteq \bot \tag{5.13}$$

Proof 5.2.2 : *Assume* $\exists y$, *that :*

$$p_1(x_1, y) \tag{5.14}$$
$$p_2(x_2, y) \tag{5.15}$$

We assert that :

$$(5.11) + (5.14) \Rightarrow C_1(y) \tag{5.16}$$
$$(5.12) + (5.15) \Rightarrow C_2(y) \tag{5.17}$$
$$(5.16) + (5.17) \Rightarrow C_1 \sqcap C_2 \not\sqsubseteq \bot \tag{5.18}$$
$$(5.13) + (5.18) \Rightarrow \bot \; absurd \tag{5.19}$$

Symmetric/asymmetric property

Symmetric property indicates that the relationship between two instances is bi-directional, even if the relationship is only declared in one direction, Sara *sisterOf* Lara as an example.

Asymmetric property means that the object property which is expressed between two instances a and b cannot be expressed between b and a, Sara *hasFather* Tim as an example.

Theorem 5.2.3 : Let p_1 & p_2 be two predicates, p_1 & p_2 cannot be synonyms if :

$$p_1 \; is \; a \; SymmetricProperty \; where \; p_1(x, y) \Rightarrow p_1(y, x) \tag{5.20}$$
$$p_2 \; is \; an \; AsymmetricProperty \; where \; p_2(x, y) \nRightarrow p_2(y, x) \tag{5.21}$$

Proof 5.2.3 : *Assume that* $p_1 \equiv p_2$, *then :*

$$p_1 \sqsubseteq p_2 \wedge p_2 \sqsubseteq p_1 \tag{5.22}$$

We assert that :

$$(5.22) \Rightarrow p_1(x,y) \Rightarrow p_2(x,y) \tag{5.23}$$
$$(5.20) \Rightarrow p_1(x,y) \Rightarrow p_2(y,x) \tag{5.24}$$
$$(5.20) + (5.22) \Rightarrow p_2(x,y) \Rightarrow p_2(y,x) \tag{5.25}$$
$$(5.21) + (5.24) \Rightarrow \perp \; absurd \tag{5.26}$$

which is impossible because p_2 *is AsymmetricProperty*

Transitive property

A property P is transitive, this means if $a\ P\ b$ and $b\ P\ c$ then $a\ P\ c$. For example, if Adam *hasNeighbor* Saly and Saly *hasNeighbor* Taylor then Adam *hasNeighbor* Taylor. Therefore, if the property is not transitive, that means the relation does not allow to bind the first individual to the last one. For example, Alice *hasFriend* Elsie and Elsie *hasFriend* Bob, so it is not necessarily that Alice *hasFriend* Bob.

Theorem 5.2.4 : Let p_1 & p_2 be two predicates, p_1 & p_2 cannot be synonyms if :

$$p_1 \; is \; a \; TransitiveProperty \; where \; p_1(x,y) \wedge p_1(y,z) \Rightarrow p_1(x,z) \tag{5.27}$$
$$p_2 \; is \; a \; Non \; TransitiveProperty \; where \; p_2(x,y) \wedge p_2(y,z) \nRightarrow p_2(x,z) \tag{5.28}$$

Proof 5.2.4 : *Assume that* $p_1 \equiv p_2$, *then :*

$$p_1 \sqsubseteq p_2 \wedge p_2 \sqsubseteq p_1 \tag{5.29}$$

We assert that $\forall x,y,z$ *:*

$$(5.29) \Rightarrow p_1(x,y) \Rightarrow p_2(x,y) \tag{5.30}$$
$$(5.29) \Rightarrow p_1(y,z) \Rightarrow p_2(y,z) \tag{5.31}$$
$$(5.29) \Rightarrow p_1(x,z) \Rightarrow p_2(x,z) \tag{5.32}$$
$$(5.27) + (5.30) + (5.31) + (5.32) \Rightarrow p_2 \; is \; a \; TransitiveProperty \tag{5.33}$$
$$(5.28) + (5.33) \Rightarrow \perp \; absurd \tag{5.34}$$

Functional property

A property P is functional, this means that it can have only one unique range value y (individuals or data values) for each instance x. For example, a person has only one biological mother, Toni *hasBoilogicalMother* Yos. On the contrary, a non-functional property can have, for the same instance x, several range values. For example, a person may have several children. Yos *hasChild* Toni and Yos *hasChild* Tara.

Theorem 5.2.5 : Let p_1 & p_2 be two predicates, p_1 & p_2 cannot be synonyms if :

$$p_1 \; is \; a \; FunctionalProperty \tag{5.35}$$
$$p_2 \; is \; a \; Non \; FunctionalProperty \tag{5.36}$$

Proof 5.2.5 : *Assume that* $p_1 \equiv p_2$, *then :*

$$p_1 \sqsubseteq p_2 \wedge p_2 \sqsubseteq p_1 \tag{5.37}$$

We assert that :

$$(5.36) \Rightarrow \exists y_1, y_2 \mid p_2(x, y_1) \wedge p_2(x, y_2) \Rightarrow y_1 \nsim y_2 \tag{5.38}$$
$$(5.37) \Rightarrow p_2(x, y_1) \Rightarrow p_1(x, y_1) \tag{5.39}$$
$$(5.37) \Rightarrow p_2(x, y_2) \Rightarrow p_1(x, y_2) \tag{5.40}$$
$$(5.35) + (5.39) + (5.40) \Rightarrow y_1 \sim y_2 \tag{5.41}$$
$$(5.38) + (5.41) \Rightarrow \perp \; absurd \tag{5.42}$$

Inverse functional property

This property is simply the opposite of the functional property. It means that it can have only one unique domain value x (individuals) for each object y.

Theorem 5.2.6 : Let p_1 & p_2 be two predicates, p_1 & p_2 cannot be synonyms if :

$$p_1 \; is \; a \; InverseFunctionalProperty \tag{5.43}$$
$$p_2 \; is \; a \; Non \; InverseFunctionalProperty \tag{5.44}$$

Proof 5.2.6 : *Assume that* $p_1 \equiv p_2$, *then :*

$$p_1 \sqsubseteq p_2 \wedge p_2 \sqsubseteq p_1 \tag{5.45}$$

We assert that $\exists x_1, x_2$ *:*

$$(5.44) \Rightarrow \; p_2(x_1, y) \wedge p_2(x_2, y) \Rightarrow x_1 \nsim x_2 \tag{5.46}$$
$$(5.45) \Rightarrow p_2(x_1, y) \Rightarrow p_1(x_1, y) \tag{5.47}$$
$$(5.45) \Rightarrow p_2(x_2, y) \Rightarrow p_1(x_2, y) \tag{5.48}$$
$$(5.43) + (5.47) + (5.48) \Rightarrow x_1 \sim x_2 \tag{5.49}$$
$$(5.46) + (5.49) \Rightarrow \perp \; absurd \tag{5.50}$$

Cardinality restrictions

OWL2 supports not only Functional property in order to retain just one unique object value, but also goes beyond that by providing the users an authority to specify restrictions to the number of objects values. In the following subsections, we will explain the repercussions of the *max* and *min* cardinality restrictions to find the synonym predicates.

— **Max cardinality restriction**

The *max* cardinality restriction allows describing a class of individuals that have at most N value for a given property P. For example, the plane A380-800 has a seating capacity of 868 passengers, so we cannot have more than 868 passengers in the same flight. However, a number less than the value of cardinality constraint is for sure accepted. Formally, we can express this constraint as following : *passengerA380-800 owl:maxCardinality "868"^^ xsd:nonNegativeInteger*. Therefore, we restrict to 868 the instantiation of the *passengerA380-800* for the same individual.

Theorem 5.2.7 : Let p_1 & p_2 be two predicates and $(a, b) \in \mathbb{N}$, p_1 & p_2 cannot be synonyms if :

$$p_1 \; has \; maxCardinality \leqslant a \tag{5.51}$$
$$p_2 \; has \; maxCardinality \leqslant b \tag{5.52}$$
$$a > b \tag{5.53}$$

Proof 5.2.7 : *Assume that* $p_1 \equiv p_2$, *then :*

$$p_1 \sqsubseteq p_2 \wedge p_2 \sqsubseteq p_1 \tag{5.54}$$

We assert that $(\forall i, j \in [1, a], \; i \neq j)$ *:*

$$(5.51) \Rightarrow p_1(x, y_1), ..., p_1(x, y_a) \Rightarrow \; y_i \nsim y_j \tag{5.55}$$

$$
\begin{aligned}
(5.51) + (5.54) \Rightarrow\ & p_1(x, y_1) \Rightarrow p_2(x, y_1) \\
& p_1(x, y_2) \Rightarrow p_2(x, y_2) \\
& \vdots \\
& \vdots \\
& p_1(x, y_b) \Rightarrow p_2(x, y_b) \\
& \vdots \\
& p_1(x, y_a) \Rightarrow p_2(x, y_a)
\end{aligned} \tag{5.56}
$$

$$(5.56) \Rightarrow |\{(p_2(x, y_1), \ldots, p_2(x, y_a)\}| = a \tag{5.57}$$

$$(5.52) + (5.57) \Rightarrow \perp \ absurd \tag{5.58}$$

— **Min cardinality restriction**
The *min* cardinality restriction allows describing a class of individuals that have at least N value for a given property P. For example, a father must have at least one child, and there is no limit for the number of children. We do not consider this restriction due to the Open World Assumption [DS06], that states that the lack of knowledge does not imply falsity. In the previous example, the fact that a father has no children is not considered as an inconsistency because it is possible that this father has a child (or children) that is merely unknown.

Phase 3 : NLP-based analysis

The returned candidates have shown that some predicates are semantically similar but non-equivalent such as *composer* and *artist*. The Domain and Range types of instances of these predicates are the same and share same features (e.g., asymmetric, non-functional). Thus, statistical and semantic analyses are not sufficient to detect that *composer* and *artist* are non-equivalent predicates. To address such issue, we have used a learning algorithm to map words or phrases from the vocabulary to vectors of numbers called "word embedding". Word embedding uses an efficient technique to vectorize the text by converting strings to numbers, where similar words have similar encodings. We transfer this technique from synonym detection in natural language processing [PSM14a, WCR+14] into the field of KGs. Word2Vec that was developed by [MSC+13] is one of the most popular techniques to learn word embeddings.

We apply Word2vec tool [5] for learning word embeddings based on a training dataset. The idea behind this tool is to assign a vector space to each unique word in the corpus where any words sharing common contexts are located close to each other in the space. Therefore, if there are two words that used about the same in the context, then these words are probably quite similar in meaning (e.g., *wrong* and *incorrect*) or are at least related (e.g., *France* and *Paris*).

Word2vec uses training algorithms to generate word vectors (embeddings) based on a dataset. As it would be extremely expensive to calculate the similarity of all predicate pairs of the dataset, we run Word2vec only on the resulting pairs from applying the statistical and semantic analyses in Phase 1 and Phase 2. Then, we apply cosine similarity [Sal89] for comparing two vectors, which is defined as follows :

$$\cos(\mathbf{t}, \mathbf{e}) = \frac{\mathbf{te}}{||\mathbf{t}||||\mathbf{e}||} = \frac{\sum_{i=1}^{n} \mathbf{t}_i \mathbf{e}_i}{\sqrt{\sum_{i=1}^{n} (\mathbf{t}_i)^2} \sqrt{\sum_{i=1}^{n} (\mathbf{e}_i)^2}}$$

Where similarity score will always be between 0.0 and 1.0.

A high similarity value indicates that two words are closely related and the maximum similarity (1.0) indicates that they are identical. This phase helps to decrease the number of false positive results, through including the candidates that have a significant similarity score and excluding them otherwise.

Finally, we have followed the metric identified by [MMB12] to assess the conciseness dimension at schema

5. https://code.google.com/archive/p/word2vec/

level. The conciseness at schema level is measured as follows :

$$\frac{number\ of\ unique\ predicates\ of\ a\ dataset}{number\ of\ predicates\ in\ a\ target\ schema}$$

.

5.2.4 Experimental Evaluation

In this section, we present two experiments performed on DBpedia and YAGO datasets in order to evaluate our approach. The metrics we have used for evaluating pair accuracy are the standard precision, recall and F-measure (harmonic mean of precision and recall) that are calculated as follows :

- True Positive (TP) : the number of discovered predicates by our approach that are actually equivalent predicates.
- False Positive (FP) : the number of discovered predicated by our approach that are actually non-equivalent predicates.
- False Negative (FN) : the number of equivalent predicates that are not discovered by our approach.
- $Precision = \frac{TP}{TP+FP}$
- $Recall = \frac{TP}{TP+FN}$
- $F-measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$

As it is obvious that eliminating equivalent predicates will improve the conciseness of the dataset, we only focus on evaluating the accuracy of our approach to identify synonym predicates.

Experimental setup

As we have already pointed out that DBpedia project extracts structured knowledge from Wikipedia, it should represent a good challenge for our approach to find equivalent predicates. There is a great chance that some of data entered by different contributors is equivalent. Unfortunately, as long as many datasets are available in LOD, DBpedia suffers a lack of expressive OWL2 features such as *functional* properties, *transitive* properties, etc. Therefore, in this case it is difficult for our approach to perform a semantic analysis. As illustrated in Table 5.8, only 30 functional properties that represent 1% have been defined. Furthermore, DBpedia neither use *max* cardinality nor *transitive* or *symmetric* properties. In addition, we have noted that 16.3% of predicates are represented without domains and 10.2% without ranges.

Table 5.8 : Features predicates of DBpedia dataset (v10-2016).

Feature	existence
Domain	83.7%
Range	89.8%
Functional properties	1%
Transitive properties	0%
Symmetric properties	0%
Max cardinality	0%

To address the lack of domains and ranges, we infer, based on the approach proposed in [TKS12b], the missed predicate domains (and/or ranges). By studying the instances that occur with each predicate which has no *rdfs:domain* value (and/or *rdfs:range* value), we have found that some of these instances may belong to different classes. In this case, only the class having a number of instances greater than a selected threshold will be defined as a domain (or range) of the predicate. In case the number of instances is smaller than the threshold, *owl:Thing* will be selected as domain (or range) value. Besides, we have

applied [TKS12b, FVS12] to enrich DBpedia ontology with the other OWL2 properties (e.g., *functional*, *transitive*, etc.).

On the other hand, due to the fact that some predicates share the same features such as *artist*, *composer* and *writer*, we have decided to use Word2vec tool, as we explained in Section 5.2.3. Our goal is to convert each predicate to a vector based on its context, and then calculate the similarity between predicate pairs. For this reason, we need a training dataset that contains all the predicate candidates. To create this dataset, we have chosen to merge data from Large Movie Review Dataset that contains 50,000 reviews from IMDb [6], and Polarity Dataset v2.0 [7] that have 2,000 movie reviews. This choice is motivated by the fact that both datasets include the majority of the candidates according to our experiments. For missed predicates or when the frequency of the predicate is very low, we have generated paragraphs from the DBpedia dataset itself and we have added them to the training dataset. Actually, we have taken into account the suggestion of Carlson et al. [CBK+10] which proposes that 10-15 examples are typically sufficient to learn the meaning of the predicate from Natural Language texts. We have generated these paragraphs through the text that exists in *rdfs:comment* of both the subject and the object of the triple, and *rdfs:label* of the predicate.

Example 5.2.1 : The English phrases about *dbo:residence* predicate is generated using the following Listing 5.2.

```
SELECT DISTINCT ?s1 ?p1 ?o1 WHERE {
  ?s dbo:residence ?o .
  ?s rdfs:comment ?s1 .
  ?o rdfs:comment ?o1 .
  dbo:residence rdfs:label ?p1 .
  FILTER (lang(?o1) = 'en')
  FILTER (lang(?s1) = 'en')
  FILTER (lang(?p1) = 'en') }
```

Listing 5.2 : Sample query returning English phrases about *dbo:residence* predicate

As a result, we have got a paragraph that contains the missed candidate to be added to our training dataset. For example, the previous query generates the following paragraph (an excerpt) : "*Lena Headey is an English actress...residence London is the capital of England and the United Kingdom...*". Adding this paragraph to the dataset will help the training process to generate a vector for the predicate *residence*, and according to the context *residence* connects between *Person* and *Place*.

First experiment

The objective of this first experiment is to show the improvement in detecting synonyms brought by the semantic analysis and NLP phases. As a reminder, the main goal of our approach is to evaluate the statistical analysis of synonyms predicates, which is the core of Abedjan and Naumann's approach. To evaluate our results, we have used a gold standard of DBpedia synonyms predicates generated by [ZGB+17]. This gold standard contains 473 true positive pairs of 10 classes in DBpedia. Compared to [ZGB+17], our approach gives, for a support threshold equals 0.01%, a slightly better F-measure (0.76 for our approach and 0.75 for [ZGB+17]). But as we have explained just before, the objective here is to show the interest of using semantic analysis and NLP. Thus, our approach would be rather complementary with [ZGB+17] instead of being a direct concurrent.

Figure 5.6 and Figure 5.7 illustrate the number of equivalent predicates pairs and the F-measure at each phase. To show the improvement that Phase 2 and 3 make, we have chosen a low support threshold value for the statistical analysis to obtain a large number of candidates. This will increase the number of true and false positive results at Phase 1, and will show how Phase 2 and Phase 3 decrease false positive

6. http://ai.stanford.edu/~amaas/data/sentiment/
7. http://www.cs.cornell.edu/people/pabo/movie-review-data/

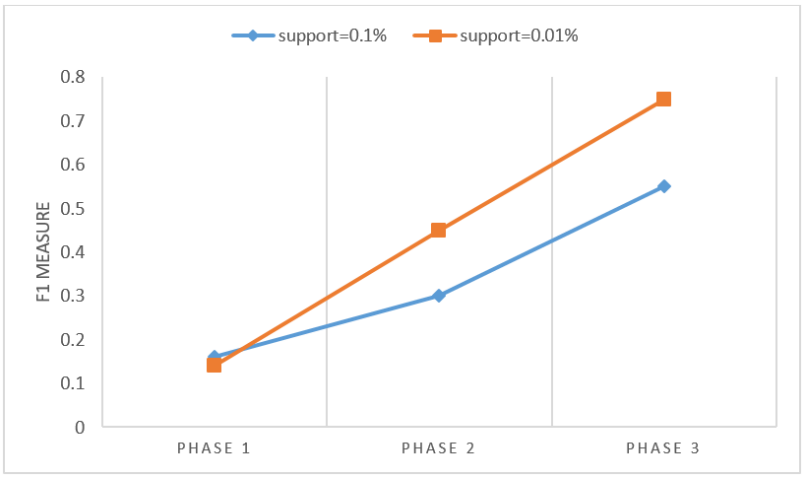

FIGURE 5.6 : F1-measure values at each phase based on support threshold.

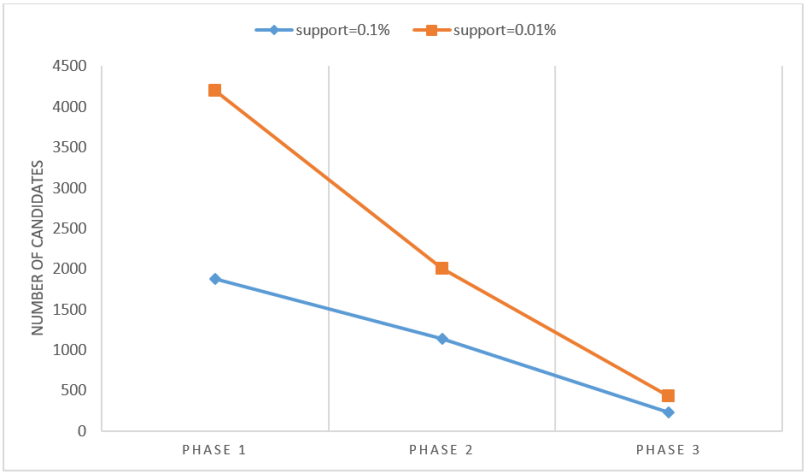

FIGURE 5.7 : Number of candidate pair at each phase.

results to enhance the precision value. Besides, at Phase 3, our approach filters the candidates regarding their similarity scores that should be less than a user-specific threshold. In this experiment, we have set the value on 50%.

Figure 5.7 shows that with a support threshold equals 0.01%, we obtain after applying the statistical analysis (Phase 1) 4197 candidate pairs. Then, by performing a semantic analysis (Phase 2), the number decreases to 2006 which represents the elimination of 52.2% of false positive results. For example, the predicates *owner* and *employer* have been proposed as equivalent predicates by the statistical analysis phase ; because on the one hand, they share a significant number of object values in their range *dbo:Organisation*, and on the other hand, they rarely co-occur for the same instance. By applying the semantic analysis, this pair of predicates will be excluded due to a domain disjointness. Indeed, the domain of *employer* is *dbo:Person* and the domain of *owner* is *dbo:Place*, and *dbo:Person* and *dbo:Agent*, that is a super class of *dbo:Person*, are disjoint. Thus, as explained in Section 5.2.3, *owner* and *employer* cannot be synonyms. Finally, by performing an NLP-based analysis (Phase 3), the number of candidate pairs decreases to 429, which represents the elimination of 78.6% of false positive results. This phase was able to filter the predicates that share the same semantic features but are non-equivalents such as *author* and *composer*.

Our approach works well to achieve our objective through decreasing the number of false positive results. The experiment shows that we can increase the precision value without affecting the recall.

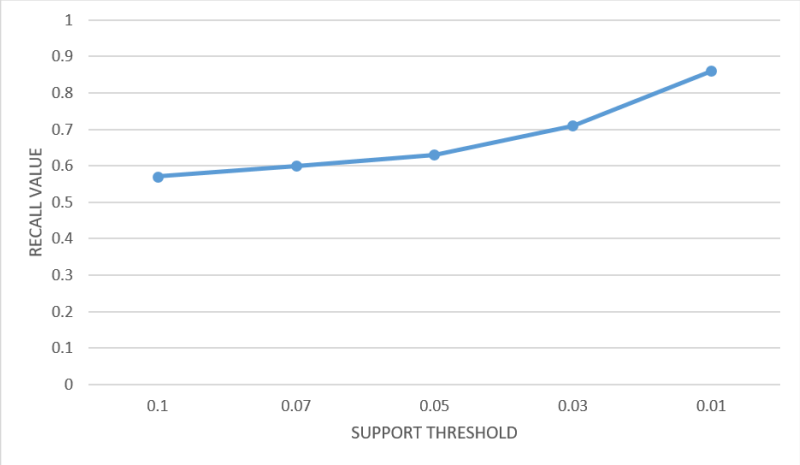

FIGURE 5.8 : Recall value based on support threshold values.

Second experiment

The gold standard of synonyms predicates of the first experiment is the only one that we have found in LOD. Thus, to perform more tests on our approach, we have created a new gold standard from mappings established between the predicates of different datasets. In fact, the mechanism consists of combining two or more datasets that have similar equivalence predicates. These predicates will therefore be the gold standard of the new dataset. For this experiment, we have chosen to merge DBpedia and YAGO that share a significant number of equivalent predicates. PARIS tool [SAS11] proposes a gold standard containing 173 mappings between the predicates of these two datasets. However, this gold standard is incomplete and, thus, can only serve to see if our approach is able to find all equivalent predicates. Due to the huge number of instances that uses the predicates of this gold standard, we have demanded from three experts to manually extract 35 mappings (equivalent predicates). Then, for each dataset, we have chosen a couple of categories from different natures to cover all of the 35 predicates. For DBpedia, we have selected instances that have the following categories : $C = \{$*dbo:Person*,*dbo:Film*, *dbo:Organisation*, *dbo:PopulatedPlace*$\}$, and for YAGO those that are semantically close to those of DBpedia : $C = \{$*yago:wordnet_person_105217688*, *yago:wordnet_movie_106613686*, *yago:wordnet_organization_1008008335*, *yago:wordnet_urban_area_108675967*$\}$. Figure 5.8 shows the obtained recall at different support thresholds of Phase 1. The maximum value is obtained when the support threshold equals 0.01%. This is logical since a great number of candidates is generated. However, certainly a lot of false positives will also be generated. The reason why we do not find all the synonyms (e.g., *isbn* and *hasISBN*) is due to the fact that some predicate pairs share insufficient number of objects. The interesting result of this experiment is that our approach finds a good number of synonyms (recall at roughly 60%) even if the support threshold is 0.1% which is relatively high in comparison to 0.01%.

5.3 Conclusion

In the first part of this chapter, we have presented an approach for revealing conceptual schemas from RDF Knowledge Graphs. The approach is an iterative process that computes a plausible model from the data values. We have shown how to automatically extract schema and represent it as a model from a data source using a user-specified threshold. The inferred model takes into account the data and user quality expectations. The result is a conceptual schema enriched by both completeness values as a relevancy indicator on the elements of the models, and existence constraints that inform about how often these elements co-exist or co-appear in the real data.The elements composing the model (classes, relationships, and properties) are obtained by applying a mining algorithm with an underlying assumption stating that the more frequent a schema is, the more relevant it is. The user can decide on the desired completeness,

the parts of the data for which the model will be inferred, and the possibility to focus on a different class through an iterative process. Currently, our demo supports only the DBpedia and Wikidata Knowledge Graphs. We have provided use cases to demonstrate the usefulness of such a tool. We believe that it can help in the discovery of a new knowledge graph and its internal structure. Therefore, it can help in the adoption of RDF Knowledge Graphs.Our analysis revealed some interesting characteristics allowing the characterization of the sources and the behavior of the community that maintains each of the data sources. The results show the rich opportunities of analysis offered by our approach and underlying outputs.

In the second part, we have proposed a new approach to evaluate the conciseness of Knowledge Graphs by discovering synonym predicates. This approach consists of three phases : (1) performing a statistical analysis to obtain an initial set of synonyms predicates, (2) performing a semantic analysis of obtained set by exploring OWL2 features (*functional, transitive, cardinality, etc.*), and (3) finally, performing a similarity-based comparison between contextual vectors representing each candidate predicate. The main objective of the last two phases is to reduce the false positive candidates generated in the first phase. We have evaluated our approach on DBpedia and YAGO datasets. The experiment results show that our approach is highly promising, as it allows eliminating about 78.6% of false positives compared to Abedjan and Naumann's approach [AN13]. They also show good results in terms of F-measure. However, it has failed to discover some synonym predicates that are rarely used. This is due to the fact that in a user defined content environments, data is not equally defined. Some categories attract more publication efforts than others. The *Film* and *Scientist* categories have richer descriptions, in terms of the number of properties, compared to *Organisation* and this phenomenon is enhanced in DBpedia too.

Conclusion and Research Perspectives

6.1 Summary

Emerged RDF Knowledge Graphs undoubtedly represent a valuable source of information that could be exploited by human and automatic agents. However, the enrichment of these data sources, which represents a crucial process for its promotion, may lead to a serious deterioration of their quality if unsuitable methods and tools are used. A thorough examination of data before adding them to KGs is essential to ensure keeping at least the same level of quality. Unfortunately, in the semantic web filed, existing methods and tools for interlinking, enriching, merging or refining data are often intended to handle generic data. Thus, it is essential to provide fine-grained solutions that consider the context (domains, categories, types, etc.) of data to be processed. In this manuscript, I mainly stressed the importance of considering the context in enriching KGs with contextual identity links and specific-domain data. Although the approaches and solution I presented are a good start, there is still much to be explored in this research direction.

In the last eight years, my contributions to the Knowledge Graph enrichment field have mainly been oriented to the following research topics

Specific-domain interlinking. In this topic, the objective was to adapt the interlinking process to the nature of data, in particular the geographic data. To this aim, we proposed a model for the representation of geometry characteristics. The first step of our work consisted in identifying the characteristics of geometries and formalizing them within a vocabulary that we called XY semantics. We relied on existing standards and works for the representation of quality, the level of detail and geometric modeling of geographic data. The particularity of this vocabulary is that it allows to associate to each spatial reference of a resource the metadata that describes its characteristics. This makes it possible to apprehend the meaning of each spatial reference, and thus manage the problem of their heterogeneity within a data source. Then we proposed an approach to solve the problem of parameterizing the comparison of spatial references in an interconnection process, when they present a strong heterogeneity within their original source and a high spatial density. This approach allows to dynamically adapt, at the level of each comparison of spatial references pair, the parameters for calculating the similarity between them. This adaptation is performed using a rule base that relies on the characteristics of the two spatial references to compare, to infer the parameters for calculating similarity. This rule base must be defined by a domain expert. Our approach overcomes the major challenge of defining the interconnection parameters in the case of a strong heterogeneity of geometries within and between datasets. We have ensured that the implementation of this approach is carried out in a generic interconnection tool so that it can be used for as many application cases as possible.

Contextual identity. to address "contextless" identity links, we proposed an approach based on sentence embedding to computes propagable properties for a given identity context. The objective is to enrich KGs with contextual identity and allow rewriting SPARQL queries to deliver more results to the user. We hypothesize that, from a semantic point of view, the closer a property is to the identity context, the more likely it could be a right candidate for propagation. Therefore, we applied sentence embedding to properties descriptions that give us numerical vectors which distributions in the vector space comply with the semantic similarity of the sentences. Our approach has been validated through quantitative and qualitative experiments. We made available on an open-source platform a toolkit for researchers and developers to explore and extend our solution.

Representing and reasoning about imprecise temporal data. in this topic we proposed an ontology-based solution that uses a fuzzy approach to represent and reason about imprecise time intervals. Hence, we extended the 4D-fluent approach with new fuzzy components to represent imprecise time intervals and the fuzzy qualitative relationships that may exist between them in the context of a fuzzy ontology. Then we extend Allen's algebra to propose fuzzy, gradual, and customized fuzzy time relations that hold between two imprecise time intervals. The definition of these relations is also based on an extension of the Vilain and Kautz point algebra which proposes a set of fuzzy, gradual and customized comparators. We proposed for each of these relations, a Mamdani IF-THEN rule.This sophisticated approach has the advantage of

being able to use a graduation of qualitative temporal relations. Our approach has been studied in the context of the VIVA and QUALHIS projects.

Knowledges Graphs quality. we were interested in two dimensions : completeness and conciseness. To assess the completeness we proposed a mining-based approach that includes two steps. The first step aims to find the properties patterns that are most shared by the subset of instances extracted from the triple store related to the same category. This set, called *transaction*, will be then used to calculate a completeness value regarding these patterns. In the second step an average is calculated to obtain the completeness of each transaction and, thus, the completeness of the whole dataset. We have implemented a web-based completeness demonstrator for DBpedia called *LOD-CM* and provided use cases to demonstrate the usefulness of such a tool.
For the conciseness, we proposed an approach to discover equivalent predicates in Knowledge Graphs. Our approach consists of three sequential phases. The first phase, which is a statistical analysis, discovers potential equivalent predicates. The next two phases exclude non-equivalent predicates based on the meaning of the predicates through the semantic features and the context where the predicate is used. Extensive experimental evaluation of real-world RDFKnowledge Graphss (DBpedia and YAGO) has been done.

6.2 Perspectives

Many research questions and perspectives are raised by our proposals. In the following, we present those that we believe are the most promising :

— The method that we proposed to dynamically adapt the parameterization of the fine-level property value comparison is applied in the case of geometries. However it can be considered for other interconnection criteria. Indeed, relying on metadata that describe the quality and conditions of acquisition of property values in order to adapt their comparison is an idea that can be used on other properties when their values are represented in a heterogeneous way within the same source. For example, the application of this approach could be useful in the case where we would like to compare properties that describe dates, knowing that the dates are formatted in a heterogeneous way. In this case, metadata that specifies the date format for each date value could be taken into account during the comparison by homogenizing the format of the two dates before calculating the distance between them.

— Regarding the approach of propagation some limitations of our approach need further investigation. Firstly, only properties with a textual description could be processed. Using other features to improve the results, like values of properties or semantic features of the property, should be tried. However, capturing ontological information of a property when embedding is still an open problem. Secondly, using only sentence embedding, combined with intuition from Tober's first law, might be naïve in some cases. Therefore, there is a need to challenge our approach with a combination of distinct knowledge graphs. For the time being, we only considered in lattices the case where the entity is subject to a triple, and we should also consider cases where it is the value of a triple. Moreover, using SPARQL queries to help the user to select the best-suited identity context might be an interesting starting point for later work. Finally, to explore SPARQL queries expansion (presented in Section 3.4.3), a prototype should be implemented to allow users to select the proper context according to the ranked list of contexts. Also, using RDF* and SPARQL* [HT14] to represent the context as defined in this chapter should be investigated.

— Concerning the representation and reasoning about imprecise time intervals : a first direction could be to define a composition table between the resulting relationships of precise and imprecise time intervals. Indeed, our temporal relationships retain the properties of Allen relationships (symmetry, transitivity and reflexivity). A second direction could be to extend our approach to represent and reason over time intervals that are both imprecise and uncertain.

— Regarding the quality assessment, we plan to investigate the role of conceptual modeling through integrated system upon several Knowledges Graphs. A first perspective could be to add more KGs

and allow the user to compare easily two conceptual schemas side by side. We believe that the ability to compare two conceptual schemas of two datasets can help to choose the one that suits better for use.

6.3 Future Research Projects

My future research projects in the short and medium term are related to the application of the Semantic Web technologies in three different domains in which I have initiated a number of collaborations : cyber and physical security, building, and musicology.

Cyber and physical security in critical infrastructures. during the past decade, many countries have faced threats and attacks that have rapidly grown. Changing the lives, habits and fears of hundreds of millions of citizens. The sources of these threats and attacks were heterogeneous. The most recent ones, both physical (theft of medical equipment) and cyber (attack of hospital information systems), are related to the Covid-19 health crisis.

Today, the boundaries between the physical and cyber worlds are increasingly blurred. Almost everything is connected to the Internet and threats cannot be categorised as physical or cyber. It is therefore necessary to develop an integrated approach in order to combat a combination of threats. In 2017, I was responsible, for the Cnam, for setting up a European project, in collaboration with various academic (ISEP, AMC, KUL) and industrial (Airbus, AP-HM, Milestone) partners. This project, that we called SAFECARE [1], and for which we obtained funding from the European Commission, deals with the issue of combining cyber and physical threats in the context of health infrastructures, which are among the most critical and vulnerable infrastructures. The objective of SAFECARE is to provide solutions that enhance physical and cyber security in a transparent and cost-effective manner. These solutions will be based on new technologies and innovative approaches to improve threat prevention, detection, incident response and impact mitigation.

Our ISID team at the Cnam CEDRIC laboratory is responsible for the development of the SAFECARE core module, which is responsible for establishing interconnections between the physical and cyber world. This module is based on an impact propagation model that formalizes the relationships between physical and cyber resources, in order to simulate the cascading propagation of effects between these different resources. On the basis of this model, it will be possible to anticipate the potential impacts of cyber and physical incidents, not only on building infrastructures (power, cooling, etc.), but also on IT systems and medical devices. We are currently working on a first version of this module using Semantic Web technologies. A preliminary ontology, which formalizes the concepts of the physical and cyber security domains and their relationships, is being implemented in the OWL 2 language. Subsequently, we plan to perform a first simulation of the propagation of impacts on a real scenario. To do so, we will use a reasoner to infer the propagation of impacts on cyber and physical resources. I plan to continue working on generalizing the solution implemented in SAFECARE, for application to any critical infrastructure. In addition, I plan to exploit the combination of Semantic Web technologies with deep learning algorithms, which have become very powerful, in order to improve risk and threat prevention. In this regard, I have initiated discussions with EADS Cassidian CyberSecurity to establish a new collaboration.

Building domain. in 2018, I started a collaboration with the Scientific and Technical Centre for Building (CSTB) and the Laboratory for Computer Science at Paris-Saclay University, on the capitalisation of knowledge, of researchers and experts in the field of building, in order to allow its sharing and reuse in other fields and for other uses. This knowledge must keep its meaning and context to allow an efficient capitalization. This initiative follows the work started in an internal project at CSTB which aims to offer a support (a tool) to identify materials containing asbestos, based on a heterogeneous corpus.

This first collaboration is part of Thamer Mecharnia's thesis that I am currently co-supervising with Lydia Chibout (CSTB) and Nathalie Pernelle (Paris-Saclay University). In this thesis, we proposed a first method for predicting the presence of asbestos, based on temporal data, describing the probability of the presence of asbestos in marketed products. To achieve our goal, we are working on the creation of an asbestos

1. `https://www.safecare-project.eu`

ontology that will be populated using data from external resources. This information will then be used to calculate the probability of asbestos in the components of a given building. At the same time, I am currently working with the CSTB on setting up an ANR project on the reduction of residential energy consumption, which has increased significantly in France and Europe in recent years. Our proposal is based on the development of a platform, combining different databases, to evaluate energy saving actions in buildings. These databases cover the most innovative renovation techniques in Europe, their respective energy savings, as well as the associated costs and the financial aid available for their implementation. The ultimate objective of this platform is to provide the information necessary for potential users to make the most optimal choice in terms of construction and/or renovation, with the aim of reducing their energy consumption and bills.

Musicology domain. in 2017, I worked with colleagues of the Vertigo team of CEDRIC Laboratory on a CNRS project (GioQoso) which deals with quality issues in old musical scores. These scores are precious heritage assets, intensively studied by musicological researchers. My role was to explore what the use of Semantic Web technologies could bring to the approaches of quality assessment of musical scores.
As a first step, we proposed a quality management approach based on a formal modeling, in the form of an ontology, of musicologists' expertise in music notation. We then used this model to express contextual rules that can be evaluated, either a priori, to avoid the production of erroneous scores, or a posteriori, to evaluate quality indicators concerning a score or a corpus of scores.
In a second step, we proposed an approach that extends the coding of scores with semantic annotations. This approach is based on an ontology of music notation, designed to integrate semantic musical elements, extracted from the score coding or produced by a knowledge extraction process. A mechanism for extracting RDF facts and evaluating semantic rules has been tested on a concrete example, based on the identification of dissonances in a Renaissance counterpoint. Currently, we are setting up a European project on the publication of music notation data on the Web in order to establish interconnections between distributed resources. We will do this by using techniques and approaches we have developed to publish, interconnect and measure the quality of music notations.

In the long term, I plan to continue studying the fully automatic adaptation of the Knowledge Graph interlinking, enrichment, refinement, and reasoning to the context of use. Indeed, as shown in the different approaches presented in this manuscript, the adaptation of the different process is essential to ensure a good level of quality of the data published in the Web of Data and, thus, promote the Semantic Web vision. For instance, concerning the identity in KGs, the two following citations of Natasha Noy et al. [NGJ+19] illustrate the importance of this challenge :

— "*While entity disambiguation and resolution is an active research area in the semantic Web, and now in knowledge graphs for several years, it is almost surprising that it continues to be one of the top challenges in the industry almost across the board.*"

— "*How can identity be described in a way that different teams can agree on it and know what the other teams are describing ?*"

These citations show two things : *(i)* knowing whether two objects are identical remains one of the most difficult and important problems, even for large companies like IBM, Microsoft, Google, eBay and Facebook, and, *(ii)* that it is difficult to get consensus among many people regarding identity. Therefore, generic approach cannot be an answer, and it is essential to take this notion of context into account in the definition and the usage of Knowledge Graphs. Given the enormous volume of data available today on the Web, exploring deep learning algorithms could constitute a promising research orientation towards the automation of the consideration of contexts in the various processes.

Bibliography

[Aba12] Nathalie Abadie. *Formalisation, acquisition et mise en œuvre de connaissances pour l'intégration virtuelle de bases de données géographiques : les spécifications au cœur du processus d'intégration*. PhD thesis, Université Paris-Est, 2012.

[ABT16] Manel Achichi, Zohra Bellahsene, and Konstantin Todorov. A survey on web data linking. *Ingénierie des Systèmes d'Information*, 21(5-6):11–29, 2016.

[ACH+91] Esther M Arkin, L Paul Chew, Daniel P Huttenlocher, Klara Kedem, and Joseph S Mitchell. An efficiently computable metric for comparing polygonal shapes. Technical report, CORNELL UNIV ITHACA NY, 1991.

[AF00] Alessandro Artale and Enrico Franconi. A survey of temporal extensions of description logics. *Ann. Math. Artif. Intell.*, 30(1-4):171–210, 2000.

[All83] James F. Allen. Maintaining knowledge about temporal intervals. *Commun. ACM*, 26(11):832–843, 1983.

[ALRG10] Benjamin Adams, Linna Li, Martin Raubal, and Michael F Goodchild. A general framework for conflation. *Extended Abstracts Volume, GIScience*, 2010.

[AN11] Ziawasch Abedjan and Felix Naumann. Context and target configurations for mining rdf data. In *Proceedings of the 1st international workshop on Search and mining entity-relationship data*, pages 23–24. ACM, 2011.

[AN13] Ziawasch Abedjan and Felix Naumann. Synonym analysis for predicate expansion. In *Extended Semantic Web Conference*, pages 140–154. Springer, 2013.

[BBDR+13] Sean Bechhofer, Iain Buchan, David De Roure, Paolo Missier, John Ainsworth, Jiten Bhagat, Philip Couch, Don Cruickshank, Mark Delderfield, Ian Dunlop, et al. Why linked data is not enough for scientists. *Future Generation Computer Systems*, 29(2):599–611, 2013.

[BBH+05] Marcel Ball, Harold Boley, David Hirtle, Jing Mei, and Bruce Spencer. Implementing ruleml using schemas, translators, and bidirectional interpreters., 2005.

[BG06] Silvana Badaloni and Massimiliano Giacomin. The algebra ia^{fuz} : a framework for qualitative fuzzy temporal reasoning. *Artif. Intell.*, 170(10):872–908, 2006.

[BGJM17] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *TACL*, 5:135–146, 2017.

[BHA99] Atef Bel Hadj Ali. Geometrical matching of polygons in giss and assessment of geometrical quality of polygons. In *Proceedings of International Symposium on Spatial Data Quality*, 1999.

[BMS97] Sergey Brin, Rajeev Motwani, and Craig Silverstein. Beyond market baskets : Generalizing association rules to correlations. *Acm Sigmod Record*, 26(2):265–276, 1997.

[BP11] Sotiris Batsakis and Euripides G. M. Petrakis. SOWL : A framework for handling spatio-temporal information in OWL 2.0. In Nick Bassiliades, Guido Governatori, and Adrian Paschke, editors, *Rule-Based Reasoning, Programming, and Applications - 5th International Symposium, RuleML 2011 - Europe, Barcelona, Spain, July 19-21, 2011. Proceedings*, volume 6826 of *Lecture Notes in Computer Science*, pages 242–249. Springer, 2011.

[BS08] Fernando Bobillo and Umberto Straccia. fuzzydl : An expressive fuzzy description logic reasoner. In *FUZZ-IEEE 2008, IEEE International Conference on Fuzzy Systems, Hong Kong, China, 1-6 June, 2008, Proceedings*, pages 923–930. IEEE, 2008.

[BS11] Fernando Bobillo and Umberto Straccia. Fuzzy ontology representation using OWL 2. *Int. J. Approx. Reasoning*, 52(7):1073–1094, 2011.

[BS16] Carlo Batini and Monica Scannapieco. *Data and Information Quality - Dimensions, Principles and Techniques*. Data-Centric Systems and Applications. Springer, 2016.

[BSvH16] Wouter Beek, Stefan Schlobach, and Frank van Harmelen. A contextualised semantics for owl : sameas. In Harald Sack, Eva Blomqvist, Mathieu d'Aquin, Chiara Ghidini, Simone Paolo Ponzetto, and Christoph Lange, editors, *The Semantic Web. Latest Advances and New Domains - 13th International Conference, ESWC 2016, Heraklion, Crete, Greece, May 29 - June 2, 2016, Proceedings*, volume 9678 of *Lecture Notes in Computer Science*, pages 405–419. Springer, 2016.

[CBK+10] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R Hruschka, and Tom M Mitchell. Toward an architecture for never-ending language learning. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.

[CG97] Scott D Cohen and Leonidas J Guibas. Partial matching of planar polylines under similarity transformations. In *8th Annual ACMSIAM Symposium on Discrete Algorithms*, pages 777–786, 1997.

[CKS+17] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. In *EMNLP*, pages 670–680. Association for Computational Linguistics, 2017.

[Cos14] Benoît Costes. Matching old hydrographic vector data from cassini's maps. *e-Perimetron*, 9(2):51–65, 2014.

[CPF15] Klitos Christodoulou, Norman W. Paton, and Alvaro A. A. Fernandes. Structure inference for linked data sources using clustering. *Trans. Large-Scale Data- and Knowledge-Centered Systems*, 19:1–25, 2015.

[CYK+18] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Universal sentence encoder. *CoRR*, abs/1803.11175, 2018.

[DBC08] Fabien Duchateau, Zohra Bellahsene, and Remi Coletta. A flexible approach for planning schema matching algorithms. *On the Move to Meaningful Internet Systems : OTM 2008*, pages 249–264, 2008.

[DNPR13] Fariz Darari, Werner Nutt, Giuseppe Pirrò, and Simon Razniewski. Completeness statements about RDF data sources and their use for query answering. In Harith Alani, Lalana Kagal, Achille Fokoue, Paul T. Groth, Chris Biemann, Josiane Xavier Parreira, Lora Aroyo, Natasha F. Noy, Chris Welty, and Krzysztof Janowicz, editors, *The Semantic Web - ISWC 2013 - 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part I*, volume 8218 of *Lecture Notes in Computer Science*, pages 66–83. Springer, 2013.

[DP89] Didier Dubois and Henri Prade. Processing fuzzy temporal knowledge. *IEEE Trans. Systems, Man, and Cybernetics*, 19(4):729–744, 1989.

[DPS98] Thomas Devogele, Christine Parent, and Stefano Spaccapietra. On spatial database integration. *International Journal of Geographical Information Science*, 12(4):335–352, 1998.

[DS06] Nick Drummond and Rob Shearer. The open world assumption. In *eSI Workshop : The Closed World of Databases meets the Open World of the Semantic Web*, volume 15, 2006.

[DSFM10] Li Ding, Joshua Shinavier, Tim Finin, and Deborah L McGuinness. owl : sameas and linked data : An empirical study. In *Proceedings of the Second Web Science Conference*, 2010.

[EL13] David W. Embley and Stephen W. Liddle. Big data - conceptual modeling to the rescue. In *Conceptual Modeling - 32th International Conference, ER 2013, Hong-Kong, China, November 11-13, 2013. Proceedings*, pages 1–8, 2013.

[EW16] Lisa Ehrlinger and Wolfram Wöß. Towards a definition of knowledge graphs. In Michael Martin, Martí Cuquet, and Erwin Folmer, editors, *Joint Proceedings of the Posters and Demos Track of the 12th International Conference on Semantic Systems - SEMANTiCS2016 and the 1st International Workshop on Semantic Change & Evolving Semantics (SuCCESS'16) co-located with the 12th International Conference on Semantic Systems (SEMANTiCS 2016), Leipzig, Germany, September 12-15, 2016*, volume 1695 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2016.

[FAH14] Abdelfettah Feliachi, Nathalie Abadie, and Fayçal Hamdi. Intégration et visualisation de données liées thématiques sur un référentiel géographique. In Chantal Reynaud, Arnaud Martin, and René Quiniou, editors, *14èmes Journées Francophones Extraction et Gestion des Connaissances, EGC 2014, Rennes, France, 28-32 Janvier, 2014*, volume E-26 of *Revue des Nouvelles Technologies de l'Information*, pages 35–46. Hermann-Éditions, 2014.

[FAH17a] Abdelfettah Feliachi, Nathalie Abadie, and Fayçal Hamdi. An adaptive approach for interlinking georeferenced data. In Óscar Corcho, Krzysztof Janowicz, Giuseppe Rizzo, Ilaria Tiddi, and Daniel Garijo, editors, *Proceedings of the Knowledge Capture Conference, K-CAP 2017, Austin, TX, USA, December 4-6, 2017*, pages 12:1–12:8. ACM, 2017.

[FAH17b] Abdelfettah Feliachi, Nathalie Abadie, and Fayçal Hamdi. Assessing the positional planimetric accuracy of dbpedia georeferenced resources. In Sergio de Cesare and Ulrich Frank, editors, *Advances in Conceptual Modeling - ER 2017 Workshops AHA, MoBiD, MREBA, OntoCom, and QMMQ, Valencia, Spain, November 6-9, 2017, Proceedings*, volume 10651 of *Lecture Notes in Computer Science*, pages 227–237. Springer, 2017.

[FAHA13] Abdelfettah Feliachi, Nathalie Abadie, Fayçal Hamdi, and Ghislain Auguste Atemezing. Interlinking and visualizing linked open data with geospatial reference data. In Pavel Shvaiko, Jérôme Euzenat, Kavitha Srinivas, Ming Mao, and Ernesto Jiménez-Ruiz, editors, *Proceedings of the 8th International Workshop on Ontology Matching co-located with the 12th International Semantic Web Conference (ISWC 2013), Sydney, Australia, October 21, 2013*, volume 1111 of *CEUR Workshop Proceedings*, pages 237–238. CEUR-WS.org, 2013.

[FES14] Zhengjie Fan, Jérôme Euzenat, and François Scharffe. Learning concise pattern for interlinking with extended version space. In *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences on*, volume 1, pages 70–77. IEEE, 2014.

[FH11] Christian Fürber and Martin Hepp. Swiqa-a semantic web information quality assessment framework. In *ECIS*, volume 15, page 19, 2011.

[FMG+13] Javier D. Fernández, Miguel A. Martínez-Prieto, Claudio Gutiérrez, Axel Polleres, and Mario Arias. Binary RDF representation for publication and exchange (HDT). *J. Web Semant.*, 19:22–41, 2013.

[FNS11] Alfio Ferrara, Andriy Nikolov, and François Scharffe. Data linking for the semantic web. *Int. J. Semantic Web Inf. Syst.*, 7(3):46–76, 2011.

[FNS13] Alfio Ferraram, 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.

[Fre92] Christian Freksa. Temporal reasoning based on semi-intervals. *Artif. Intell.*, 54(1):199–227, 1992.

[FVS12] Daniel Fleischhacker, Johanna Völker, and Heiner Stuckenschmidt. Mining rdf data for property axioms. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pages 718–735. Springer, 2012.

[FWJY12] Linyun Fu, Haofen Wang, Wei Jin, and Yong Yu. Towards better understanding and utilizing relations in dbpedia. *Web Intelligence and Agent Systems : An International Journal*, 10(3):291–303, 2012.

[GHM19a] Fatma Ghorbel, Fayçal Hamdi, and Elisabeth Métais. Ontology-based representation and reasoning about precise and imprecise time intervals. In *2019 IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2019, New Orleans, LA, USA, June 23-26, 2019*, pages 1–8. IEEE, 2019.

[GHM+19b] Fatma Ghorbel, Fayçal Hamdi, Elisabeth Métais, Nebrasse Ellouze, and Faïez Gargouri. Ontology-based representation and reasoning about precise and imprecise temporal data : A fuzzy-based view. *Data Knowl. Eng.*, 124, 2019.

[GHM20] Fatma Ghorbel, Fayçal Hamdi, and Elisabeth Métais. Dealing with precise and imprecise temporal data in crisp ontology. *IJITWE*, 15(2):30–49, 2020.

[GHP94] Hans Werner Guesgen, Joachim Hertzberg, and Anne Philpott. Towards implementing fuzzy allen relations. In *Proceedings of the ECAI-94 Workshop on Spatial and Temporal Reasoning*, pages 49–55, 1994.

[GHY17] Aymen Gammoudi, Allel Hadjali, and Boutheina Ben Yaghlane. Fuzz-time : an intelligent system for managing fuzzy temporal information. *Int. J. Intelligent Computing and Cybernetics*, 10(2):200–222, 2017.

[Gir12] JF Girres. Modèle d'estimation de l'imprécision des mesures géométriques de données géographiques. application aux mesures de longueur et de surface. *PhD, Université Paris–Est, France*, 2012.

[GTHS15] Luis Galárraga, Christina Teflioudi, Katja Hose, and Fabian M Suchanek. Fast rule mining in ontological knowledge bases with amie+. *The VLDB Journal—The International Journal on Very Large Data Bases*, 24(6):707–730, 2015.

[GTJ+13] Kalpa Gunaratna, Krishnaprasad Thirunarayan, Prateek Jain, Amit Sheth, and Sanjaya Wijeratne. A statistical and schema independent approach to identify equivalent properties on linked data. In *Proceedings of the 9th International Conference on Semantic Systems*, pages 33–40. ACM, 2013.

[GW02] Nicola Guarino and Christopher A. Welty. Evaluating ontological decisions with ontoclean. *Commun. ACM*, 45(2):61–65, 2002.

[GZ01] Karam Gouda and Mohammed Javeed Zaki. Efficiently mining maximal frequent itemsets. In Nick Cercone, Tsau Young Lin, and Xindong Wu, editors, *Proceedings of the 2001 IEEE International Conference on Data Mining, 29 November - 2 December 2001, San Jose, California, USA*, pages 163–170. IEEE Computer Society, 2001.

[GZ03] Gösta Grahne and Jianfei Zhu. Efficiently using prefix-trees in mining frequent itemsets. In *Proceedings of the ICDM 2003 Workshop on Frequent Itemset Mining Implementations, 19 December 2003, Melbourne, Florida, USA*, 2003.

[Ham11] Fayçal Hamdi. *Améliorer l'interopérabilité sémantique : applicabilité et utilité de l'alignement d'ontologies. (Enhancing the semantic interoperability : applicability and utility of the ontology alignment)*. PhD thesis, University of Paris-Sud, Orsay, France, 2011.

[HB10] S Hahmann and D Burghardt. Connecting linkedgeodata and geonames in the spatial semantic web. In *6th International GIScience Conference*, 2010.

[HHM+10] 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 Peter F. Patel-Schneider, Yue Pan, Pascal Hitzler, Peter Mika, Lei Zhang, Jeff Z. Pan, Ian Horrocks, and Birte Glimm, editors, *The Semantic Web - ISWC 2010 - 9th International Semantic Web Conference, ISWC 2010, Shanghai, China, November 7-11, 2010, Revised Selected Papers, Part I*, volume 6496 of *Lecture Notes in Computer Science*, pages 305–320. Springer, 2010.

[HHM+15] Noura Herradi, Fayçal Hamdi, Elisabeth Métais, Fatma Ghorbel, and Assia Soukane. Personlink : An ontology representing family relationships for the CAPTAIN MEMO memory prosthesis. In Manfred A. Jeusfeld and Kamalakar Karlapalem, editors, *Advances in Conceptual Modeling - ER 2015 Workshops, AHA, CMS, EMoV, MoBiD, MORE-BI, MReBA, QMMQ, and SCME Stockholm, Sweden, October 19-22, 2015, Proceedings*, volume 9382 of *Lecture Notes in Computer Science*, pages 3–13. Springer, 2015.

[HHT07] Mirella Huza, Mounira Harzallah, and Francky Trichet. Ontomas : a tutoring system dedicated to ontology matching. In *Enterprise Interoperability II*, pages 377–388. Springer, 2007.

[HKS06] Ian Horrocks, Oliver Kutz, and Ulrike Sattler. The even more irresistible sroiq. *Kr*, 6:57–67, 2006.

[HPSB+04] Ian Horrocks, Peter F. Patel-Schneider, Harold Boley, Said Tabet, Benjamin Grosofand, and Mike Dean. SWRL : A semantic web rule language combining OWL and RuleML. W3C Member Submission, 2004.

[HPY00] Jiawei Han, Jian Pei, and Yiwen Yin. Mining frequent patterns without candidate generation. In Weidong Chen, Jeffrey F. Naughton, and Philip A. Bernstein, editors, *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, May 16-18, 2000, Dallas, Texas, USA*, pages 1–12. ACM, 2000.

[HPYM04] Jiawei Han, Jian Pei, Yiwen Yin, and Runying Mao. Mining frequent patterns without candidate generation : A frequent-pattern tree approach. *Data mining and knowledge discovery*, 8(1):53–87, 2004.

[HT14] Olaf Hartig and Bryan Thompson. Foundations of an alternative approach to reification in RDF. *CoRR*, abs/1406.3399, 2014.

[IHC19] Subhi Issa, Fayçal Hamdi, and Samira Si-Said Cherfi. Enhancing the conciseness of linked data by discovering synonym predicates. In Christos Douligeris, Dimitris Karagiannis, and Dimitris Apostolou, editors, *Knowledge Science, Engineering and Management - 12th International Conference, KSEM 2019, Athens, Greece, August 28-30, 2019, Proceedings, Part I*, volume 11775 of *Lecture Notes in Computer Science*, pages 739–750. Springer, 2019.

[IHvH+17] Al Koudous Idrissou, Rinke Hoekstra, Frank van Harmelen, Ali Khalili, and Peter van den Besselaar. Is my : sameas the same as your : sameas ?: Lenticular lenses for context-specific identity. In Óscar Corcho, Krzysztof Janowicz, Giuseppe Rizzo, Ilaria Tiddi, and Daniel Garijo, editors, *Proceedings of the Knowledge Capture Conference, K-CAP 2017, Austin, TX, USA, December 4-6, 2017*, pages 23:1–23:8. ACM, 2017.

[IJB11] Robert Isele, Anja Jentzsch, and Christian Bizer. Efficient multidimensional blocking for link discovery without losing recall. In *WebDB*, 2011.

[IPH17] Subhi Issa, Pierre-Henri Paris, and Fayçal Hamdi. Assessing the completeness evolution of DBpedia : A case study. In Sergio de Cesare and Ulrich Frank, editors, *Advances in Conceptual Modeling - ER 2017 Workshops AHA, MoBiD, MREBA, OntoCom, and QMMQ, Valencia, Spain, November 6-9, 2017, Proceedings*, volume 10651 of *Lecture Notes in Computer Science*, pages 238–247. Springer, 2017.

[IPHC19a] Subhi Issa, Pierre-Henri Paris, Fayçal Hamdi, and Samira Si-Said Cherfi. Revealing the conceptual schemas of RDF datasets. In Paolo Giorgini and Barbara Weber, editors, *Advanced Information Systems Engineering - 31st International Conference, CAiSE 2019, Rome, Italy, June 3-7, 2019, Proceedings*, volume 11483 of *Lecture Notes in Computer Science*, pages 312–327. Springer, 2019.

[IPHC19b] Subhi Issa, Pierre-Henri Paris, Fayçal Hamdi, and Samira Si-Said Cherfi. Revealing the conceptual schemas of rdf datasets. In *International Conference on Advanced Information Systems Engineering*, pages 312–327. Springer, 2019.

[ISO13] ISO. 19157 : Geographic information – data quality. International standard, International Organization for Standardization (http://www.iso.org), 2013.

[Jac08] Paul Jaccard. Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaud. Sci. Nat.*, 44:223–270, 1908.

[JHY+10] Prateek Jain, Pascal Hitzler, Peter Z. Yeh, Kunal Verma, and Amit P. Sheth. Linked data is merely more data. In *Linked Data Meets Artificial Intelligence, Papers from the 2010 AAAI Spring Symposium, Technical Report SS-10-07, Stanford, California, USA, March 22-24, 2010*. AAAI, 2010.

[Jr.98] Roberto J. Bayardo Jr. Efficiently mining long patterns from databases. In Laura M. Haas and Ashutosh Tiwary, editors, *SIGMOD 1998, Proceedings ACM SIGMOD International Conference on Management of Data, June 2-4, 1998, Seattle, Washington, USA.*, pages 85–93. ACM Press, 1998.

[JRK11] Krzysztof Janowicz, Martin Raubal, and Werner Kuhn. The semantics of similarity in geographic information retrieval. *Journal of Spatial Information Science*, 2011(2):29–57, 2011.

[KC04] Graham Klyne and Jeremy J Carroll. Resource description framework (rdf) : Concepts and abstract syntax. *http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/*, 2004.

[KF01] Michel C. A. Klein and Dieter Fensel. Ontology versioning on the semantic web. In Isabel F. Cruz, Stefan Decker, Jérôme Euzenat, and Deborah L. McGuinness, editors, *Proceedings of SWWS'01, The first Semantic Web Working Symposium, Stanford University, California, USA, July 30 - August 1, 2001*, pages 75–91, 2001.

[KK07] Marinos Kavouras and Margarita Kokla. *Theories of geographic concepts : ontological approaches to semantic integration*. CRC Press, 2007.

[KK15] Kenza Kellou-Menouer and Zoubida Kedad. Schema discovery in RDF data sources. In Paul Johannesson, Mong-Li Lee, Stephen W. Liddle, Andreas L. Opdahl, and Oscar Pastor López, editors, *Conceptual Modeling - 34th International Conference, ER 2015, Stockholm, Sweden, October 19-22, 2015, Proceedings*, volume 9381 of *Lecture Notes in Computer Science*, pages 481–495. Springer, 2015.

[KS03] Yannis Kalfoglou and Marco Schorlemmer. Ontology mapping : the state of the art. *The knowledge engineering review*, 18(1):1–31, 2003.

[LM09] Justin J. Levandoski and Mohamed F. Mokbel. RDF data-centric storage. In *IEEE International Conference on Web Services, ICWS 2009, Los Angeles, CA, USA, 6-10 July 2009*, pages 911–918. IEEE Computer Society, 2009.

[LP15] Roman Lukyanenko and Jeffrey Parsons. Principles for modeling user-generated content. In Paul Johannesson, Mong-Li Lee, Stephen W. Liddle, Andreas L. Opdahl, and Oscar Pastor López, editors, *Conceptual Modeling - 34th International Conference, ER 2015, Stockholm, Sweden, October 19-22, 2015, Proceedings*, volume 9381 of *Lecture Notes in Computer Science*, pages 432–440. Springer, 2015.

[LPS19] Roman Lukyanenko, Jeffrey Parsons, and Binny M. Samuel. Representing instances : the case for reengineering conceptual modelling grammars. *Eur. J. Inf. Syst.*, 28(1):68–90, 2019.

[LSKW02] Yang W Lee, Diane M Strong, Beverly K Kahn, and Richard Y Wang. Aimq : a methodology for information quality assessment. *Information & management*, 40(2):133–146, 2002.

[LSL+06] Yuangui Lei, Marta Sabou, Vanessa Lopez, Jianhan Zhu, Victoria Uren, and Enrico Motta. An infrastructure for acquiring high quality semantic metadata. In *European Semantic Web Conference*, pages 230–244. Springer, 2006.

[LUM07] Yuangui Lei, Victoria Uren, and Enrico Motta. A framework for evaluating semantic metadata. In *Proceedings of the 4th international conference on Knowledge capture*, pages 135–142. ACM, 2007.

[MCCD13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013.

[MD08] Sébastien Mustière and Thomas Devogele. Matching networks with different levels of detail. *GeoInformatica*, 12(4):435–453, 2008.

[MGF12] Miguel A. Martínez-Prieto, Mario Arias Gallego, and Javier D. Fernández. Exchange and consumption of huge RDF data. In Elena Simperl, Philipp Cimiano, Axel Polleres, Óscar Corcho, and Valentina Presutti, editors, *The Semantic Web : Research and Applications - 9th Extended Semantic Web Conference, ESWC 2012, Heraklion, Crete, Greece, May 27-31, 2012. Proceedings*, volume 7295 of *Lecture Notes in Computer Science*, pages 437–452. Springer, 2012.

[MGH+15] Elisabeth Métais, Fatma Ghorbel, Noura Herradi, Fayçal Hamdi, Nadira Lammari, Didier Nakache, Nebrasse Ellouze, Faiez Gargouri, and Assia Soukane. Memory prosthesis. *Non-pharmacological Therapies in Dementia*, 3(2):177–180, 2015.

[MGH+18] Elisabeth Métais, Fatma Ghorbel, Fayçal Hamdi, Nebrasse Ellouze, Noura Herradi, and Assia Soukane. Representing imprecise time intervals in OWL 2. *Enterp. Model. Inf. Syst. Archit. Int. J. Concept. Model.*, 13(Special):120–132, 2018.

[Mik05] Peter Mika. Flink : Semantic web technology for the extraction and analysis of social networks. *Web Semantics : Science, Services and Agents on the World Wide Web*, 3(2):211–223, 2005.

[MJ08] Malgorzata Mochol and Anja Jentzsch. Towards a rule-based matcher selection. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 109–119. Springer, 2008.

[MMB12] Pablo N Mendes, Hannes Mühleisen, and Christian Bizer. Sieve : linked data quality assessment and fusion. In *Proceedings of the 2012 Joint EDBT/ICDT Workshops*, pages 116–123. ACM, 2012.

[MSC+13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[NGJ+19] Natalya Fridman Noy, Yuqing Gao, Anshu Jain, Anant Narayanan, Alan Patterson, and Jamie Taylor. Industry-scale knowledge graphs : lessons and challenges. *Commun. ACM*, 62(8):36–43, 2019.

[NHNR17] 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.

[NM03] Gábor Nagypál and Boris Motik. A fuzzy model for representing uncertain, subjective, and vague temporal knowledge in ontologies. In Robert Meersman, Zahir Tari, and Douglas C. Schmidt, editors, *On The Move to Meaningful Internet Systems 2003 : CoopIS, DOA, and ODBASE - OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003*, volume 2888 of *Lecture Notes in Computer Science*, pages 906–923. Springer, 2003.

[NR06] Natasha Noy and Alan Rector. Defining n-ary relations on the semantic web. W3C note, W3C, apr 2006. http://www.w3.org/TR/2006/NOTE-swbp-n-aryRelations-20060412/.

[NUMDR08] Andriy Nikolov, Victoria Uren, Enrico Motta, and Anne N De Roeck. Integration of semantically annotated data by the knofuss architecture. In *EKAW*, pages 265–274. Springer, 2008.

[Ohl04] Hans Jürgen Ohlbach. Relations between fuzzy time intervals. In *11th International Symposium on Temporal Representation and Reasoning (TIME 2004), 1-3 July 2004, Tatihou Island, Normandie, France*, pages 44–51. IEEE Computer Society, 2004.

[Oli07] Antoni Olivé. *Conceptual modeling of information systems*. Springer, 2007.

[PHC19a] Pierre-Henri Paris, Fayçal Hamdi, and Samira Si-Said Cherfi. Interlinking rdf-based datasets : A structure-based approach. In Imre J. Rudas, János Csirik, Carlos Toro, János Botzheim, Robert J. Howlett, and Lakhmi C. Jain, editors, *Knowledge-Based and Intelligent Information & Engineering Systems : Proceedings of the 23rd International Conference KES-2019, Budapest, Hungary, 4-6 September 2019*, volume 159 of *Procedia Computer Science*, pages 162–171. Elsevier, 2019.

[PHC19b] Pierre-Henri Paris, Fayçal Hamdi, and Samira Si-Said Cherfi. A study about the use of OWL 2 semantics in RDF-based knowledge graphs. In *The Semantic Web : ESWC 2020 Satellite Events - ESWC 2020 Satellite Events*, 2019.

[PHC20a] Pierre-Henri Paris, Fayçal Hamdi, and Samira Si-said Cherfi. État des lieux de l'utilisation de OWL 2 : Analyse et proposition pour capturer les utilisations de la sémantique OWL 2 dans les graphes de connaissances RDF. *Revue des Nouvelles Technologies de l'Information*, Extraction et Gestion des Connaissances , RNTI-E-36:145–156, 2020.

[PHC20b] Pierre-Henri Paris, Fayçal Hamdi, and Samira Si-Said Cherfi. OntoSemStats : an ontology to express the use of semantics in RDF-based knowledge graphs. In *Web Engineering - 20th International Conference, ICWE 2020, Helsinki, Finland, June 9-12, 2020, Proceedings*, 2020.

[PHC20c] Pierre-Henri Paris, Fayçal Hamdi, and Samira Si-Said Cherfi. Propagation contextuelle des propriétés pour les graphes de connaissances : une approche fondée sur les plongements de phrases. In *Ingénierie des Connaissances*, 2020.

[PPEB15] Minh-Duc Pham, Linnea Passing, Orri Erling, and Peter A. Boncz. Deriving an emergent relational schema from RDF data. In Aldo Gangemi, Stefano Leonardi, and Alessandro Panconesi, editors, *Proceedings of the 24th International Conference on World Wide Web, WWW 2015, Florence, Italy, May 18-22, 2015*, pages 864–874. ACM, 2015.

[PRL+03] George Percivall, Carl Reed, Lew Leinenweber, Chris Tucker, and Tina Cary. Ogc reference model. *Open Geospatial Consortium Inc*, pages 1–108, 2003.

[PSM14a] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove : Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[PSM14b] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove : Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL, 2014.

[RB01] Erhard Rahm and Philip A Bernstein. A survey of approaches to automatic schema matching. *the VLDB Journal*, 10(4):334–350, 2001.

[RMR15] Ana-Maria Olteanu Raimond, Sébastien Mustière, and Anne Ruas. Knowledge formalization for vector data matching using belief theory. *J. Spatial Inf. Sci.*, 10(1):21–46, 2015.

[RP00] Colette Rolland and Naveen Prakash. From conceptual modelling to requirements engineering. *Ann. Software Eng.*, 10:151–176, 2000.

[RP16] Petar Ristoski and Heiko Paulheim. Rdf2vec : RDF graph embeddings for data mining. In Paul T. Groth, Elena Simperl, Alasdair J. G. Gray, Marta Sabou, Markus Krötzsch, Freddy Lécué, Fabian Flöck, and Yolanda Gil, editors, *The Semantic Web - ISWC 2016 - 15th International Semantic Web Conference, Kobe, Japan, October 17-21, 2016, Proceedings, Part I*, volume 9981 of *Lecture Notes in Computer Science*, pages 498–514, 2016.

[RPS17] Joe Raad, Nathalie Pernelle, and Fatiha Saïs. Detection of contextual identity links in a knowledge base. In Óscar Corcho, Krzysztof Janowicz, Giuseppe Rizzo, Ilaria Tiddi, and Daniel Garijo, editors, *Proceedings of the Knowledge Capture Conference, K-CAP 2017, Austin, TX, USA, December 4-6, 2017*, pages 8:1–8:8. ACM, 2017.

[Sal89] Gerard Salton. *Automatic Text Processing : The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, 1989.

[Sar07] LT Sarjakoski. Conceptual models of generalisation and multiple representation. *Generalisation of geographic information : cartographic modelling and applications*, pages 11–35, 2007.

[SAS11] Fabian M Suchanek, Serge Abiteboul, and Pierre Senellart. Paris : Probabilistic alignment of relations, instances, and schema. *Proceedings of the VLDB Endowment*, 5(3):157–168, 2011.

[SC08] Steven Schockaert and Martine De Cock. Temporal reasoning about fuzzy intervals. *Artificial Intelligence*, 172(8):1158–1193, 2008.

[Sch10] S Schade. Computer-tractable translation of geospatial data. *International Journal of Spatial Data Infrastructures Research, Revue en ligne publiée par le Joint Research Centre (European Commission)*, 5, 2010.

[SE13] Pavel Shvaiko and Jérôme Euzenat. Ontology matching : state of the art and future challenges. *IEEE Transactions on knowledge and data engineering*, 25(1):158–176, 2013.

[SH11] J Salas and Andreas Harth. Finding spatial equivalences accross multiple rdf datasets. In *Proceedings of the Terra Cognita Workshop on Foundations, Technologies and Applications of the Geospatial Web*, pages 114–126, 2011.

[Sin01] Amit Singhal. Modern information retrieval : A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43, 2001.

[SN15] Mohamed Ahmed Sherif and Axel-Cyrille Ngonga Ngomo. A systematic survey of point set distance measures for link discovery. *Semantic Web Journal.(Cited on page 18.)*, 2015.

[SNA+15] Md Seddiqui, Rudra Pratap Deb Nath, Masaki Aono, et al. An efficient metric of automatic weight generation for properties in instance matching technique. *arXiv preprint arXiv:1502.03556*, 2015.

[SPG+07] Evren Sirin, Bijan Parsia, Bernardo Cuenca Grau, Aditya Kalyanpur, and Yarden Katz. Pellet : A practical OWL-DL reasoner. *J. Web Sem.*, 5(2):51–53, 2007.

[STBP18] Sandeep Subramanian, Adam Trischler, Yoshua Bengio, and Christopher J. Pal. Learning general purpose distributed sentence representations via large scale multi-task learning. *CoRR*, abs/1804.00079, 2018.

[TKS12a] Gerald Töpper, Magnus Knuth, and Harald Sack. Dbpedia ontology enrichment for inconsistency detection. In Valentina Presutti and Helena Sofia Pinto, editors, *I-SEMANTICS 2012 - 8th International Conference on Semantic Systems, I-SEMANTICS '12, Graz, Austria, September 5-7, 2012*, pages 33–40. ACM, 2012.

[TKS12b] Gerald Töpper, Magnus Knuth, and Harald Sack. Dbpedia ontology enrichment for inconsistency detection. In *Proceedings of the 8th International Conference on Semantic Systems*, pages 33–40. ACM, 2012.

[Tob70] Waldo R Tobler. A computer movie simulating urban growth in the detroit region. *Economic geography*, 46(sup1):234–240, 1970.

[VBSC12] Luis M Vilches-Blázquez, Víctor Saquicela, and Oscar Corcho. Interlinking geospatial information in the web of data. *Bridging the Geographic Information Sciences*, pages 119–139, 2012.

[VK86] Marc B. Vilain and Henry A. Kautz. Constraint propagation algorithms for temporal reasoning. In Tom Kehler, editor, *Proceedings of the 5th National Conference on Artificial Intelligence. Philadelphia, PA, August 11-15, 1986. Volume 1 : Science.*, pages 377–382. Morgan Kaufmann, 1986.

[VN11] Johanna Völker and Mathias Niepert. Statistical schema induction. In Grigoris Antoniou, Marko Grobelnik, Elena Paslaru Bontas Simperl, Bijan Parsia, Dimitris Plexousakis, Pieter De Leenheer, and Jeff Z. Pan, editors, *The Semantic Web : Research and Applications - 8th Extended Semantic Web Conference, ESWC 2011, Heraklion, Crete, Greece, May 29-June 2, 2011, Proceedings, Part I*, volume 6643 of *Lecture Notes in Computer Science*, pages 124–138. Springer, 2011.

[Vol06] Steffen Volz. An iterative approach for matching multiple representations of street data. *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 36(Part 2/W40):101–110, 2006.

[WCR+14] Julie Weeds, Daoud Clarke, Jeremy Reffin, David Weir, and Bill Keller. Learning to distinguish hypernyms and co-hyponyms. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics : Technical Papers*, pages 2249–2259. Dublin City University and Association for Computational Linguistics, 2014.

[WCZ+15] Yanxia Wang, Deng Chen, Zhiyuan Zhao, Fu Ren, and Qingyun Du. A back-propagation neural network-based approach for multi-represented feature matching in update propagation. *Transactions in GIS*, 19(6):964–993, 2015.

[WDLW10] Yan Wang, Xiaoyong Du, Jiaheng Lu, and Xiaofang Wang. Flextable : Using a dynamic relation model to store RDF data. In Hiroyuki Kitagawa, Yoshiharu Ishikawa, Qing Li, and Chiemi Watanabe, editors, *Database Systems for Advanced Applications, 15th International Conference, DASFAA 2010, Tsukuba, Japan, April 1-4, 2010, Proceedings, Part I*, volume 5981 of *Lecture Notes in Computer Science*, pages 580–594. Springer, 2010.

[WF99] Volker Walter and Dieter Fritsch. Matching spatial data sets : a statistical approach. *International Journal of geographical information science*, 13(5):445–473, 1999.

[WF06] Christopher A. Welty and Richard Fikes. A reusable ontology for fluents in OWL. In Brandon Bennett and Christiane Fellbaum, editors, *Formal Ontology in Information Systems, Proceedings of the Fourth International Conference, FOIS 2006, Baltimore, Maryland, USA, November 9-11, 2006*, volume 150 of *Frontiers in Artificial Intelligence and Applications*, pages 226–236. IOS Press, 2006.

[WS96] Richard Y Wang and Diane M Strong. Beyond accuracy : What data quality means to data consumers. *Journal of management information systems*, pages 5–33, 1996.

[YLZ14] Bisheng Yang, Xuechen Luan, and Yunfei Zhang. A pattern-based approach for matching nodes in heterogeneous urban road networks. *Transactions in GIS*, 18(5):718–739, 2014.

[Zad75] Lotfi A. Zadeh. The concept of a linguistic variable and its application to approximate reasoning - II. *Inf. Sci.*, 8(4):301–357, 1975.

[ZGB+17] Ziqi Zhang, Anna Lisa Gentile, Eva Blomqvist, Isabelle Augenstein, and Fabio Ciravegna. An unsupervised data-driven method to discover equivalent relations in large linked datasets. *Semantic web*, 8(2):197–223, 2017.

[ZRM+13] Amrapali Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, Sören Auer, and Pascal Hitzler. Quality assessment methodologies for linked open data. *Submitted to Semantic Web Journal*, 2013.

[ZRM+16] Amrapali Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, and Sören Auer. Quality assessment for linked data : A survey. *Semantic Web*, 7(1):63–93, 2016.

Curriculum Vitae of Fayçal Hamdi

1. Curriculum Vitae

Fayçal Hamdi
19/11/1984
Cnam Paris
CEDRIC Lab – ISID Team
2 rue Conté 75003 Paris
Tel: +33 (0) 6 23 89 12 41
E-mail: Faycal.Hamdi@cnam.fr
Web: https://cedric.cnam.fr/~hamdif/

1.1. Education

2011: **Ph.D in Computer Science**, Paris-Sud University, France
Title: Enhancing the semantic interoperability: applicability and utility of the ontology alignment
Supervisor: Chantal Reynaud
Co-supervisor: Brigitte Safar
Lab: CNRS-LRI (IASI team) and INRIA Saclay – Ile-de-France (Leo team)
Thesis defended on December 2nd, 2011

2008: **Master degree in Computer Science, Paris-Sud University, France**
Paris-Sud University, CNRS-LRI (IASI team) and INRIA Saclay

2007: **Engineer degree in computer science (ranked first)**
Skikda University, Algeria

1.2. Professional Experience

Since Sept. 2012: **Associate Professor at Cnam Paris**
ISID Team (Information and Decision Systems Engineering)
CEDRIC Lab
(*Sabbatical leave from March 2018 to February 2019*)

Nov. 2011 – Aug 2012: **Post-Doc**
COGIT Lab
IGN (National Institute of Geographic and Forest Information)

Oct. 2008 – July 2011: **Ph.D Student at Paris-Sud University**
CNRS-LRI (IASI team) and INRIA Saclay – Ile-de-France (Leo team)
Computer Science Lab (LRI)

Oct. 2010 – July 2011: **Instructor at Paris-Sud University**
Computer Science Departement at IUT d'Orsay

Oct. 2008 – Oct. 2010: **Instructor at Paris-Sud University**
Business and Administration Management Department at IUT de Sceaux

Sept. 2008 – Oct. 2008: **Research Engineer**
INRIA Saclay – Ile-de-France (Gemo team)

Mar. 2008 – Sept 2008: **Master Intership**
Title: Alignment-based Partitioning of Large-scale Ontologies
CNRS-LRI (IASI team) and INRIA Saclay – Ile-de-France (Gemo team)

2. TEACHING

I carried out my main activities as an Associate Professor since 2012 at the Computer Science Department (EPN 5) of the Conservatoire National des Arts et Métiers Paris and as an instructor at the Department of Business Management and Administration of the IUT de Sceaux (Paris-Sud 11 University) between 2008 and 2010 and at the Computer Science Department of the IUT d'Orsay (Paris-Sud 11 University) between 2010 and 2011. The audience I have addressed is varied: from computer scientists at the CNAM and the IUT d'Orsay to managers at the IUT de Sceaux, requiring a pedagogy, each time, adapted to their needs.

1.3. SUMMARY OF TEACHING ACTIVITIES

I was involved in different teaching units for different levels, where I prepared course materials, ED, and TP. These courses were given in face-to-face, but also in FOD (distance learning) via the Cnam's Moodle platform.

Since Sept. 2012: Associate Professor at CNAM Paris

- Multidimensional Databases and Data Warehousing (Master's degree)
- Data Warehousing and Data Mining (Master's degree) • Database Administration (Master's & B.Sc. degree)
- Web Information Systems (B.Sc. degree)
- Introduction to Information Systems (B.Sc. degree)
- Conception of Information Systems (B.Sc. degree)
- Web programming and e-Commerce (B.Sc. degree)
- Query Processing and Optimization (Master's & B.Sc. degree)

Oct. 2010 - Sept. 2011 Instructor at IUT d'Orsay - Paris-Sud University (64 hours per year):

- Information Systems Analysis and Design (2nd year B.Sc.)
- Operating Systems (1st year B.Sc.)

Oct. 2008 - Sept. 2010 Instructor at IUT de Sceaux - Paris-Sud University (64 hours per year):

- Information and communication tools (1st year B.Sc.)
- Computer Science for Management (1st year B.Sc.)
- Database models (1st year B.Sc.)

1.4. RESPONSIBILITIES

I have taken on various module and training responsibilities within EPN IT (EPN 5). I also took part in the setting up of the Advanced Master DeSIgeo in collaboration with the ENSG and the ENPC. In addition, I am actively involved in the supervision and tutoring of engineers and bachelors, and in the different VAE, VES, VAPP and EICnam admissions juries.

SINCE 2019: **In charge of the second year of the engineering degree -** Computer Science and Information Systems, in collaboration with AFIA

2015 – 2017: **Academic Director of the Advanced Master DeSIgeo -** Decision and Geolocated Information Systems, in collaboration with ENSG and ENPC

2013 – 2016: **Academic Director of the Master SID -** Information and Decision Systems

3. RESEARCH ACTIVITIES

The aim of this section is to present a synthetic description of my experience in computer science research. My very first experience started during the Master Research internship. Then, this experience was reinforced during the three years of thesis I did within the IASI team of the LRI and the Gemo project, which became a Leo team in 2010, common between the LRI of the University Paris-Sud and INRIA Saclay-Ile-de-France. From November 2011 to September 2012, I carried out my research activities as part of a post-doctorate in the COGIT (Conception Objet et Généralisation de l'Information Topographique) laboratory of the IGN (l'institut national de l'information géographique et forestière). Currently, I am continuing my research activities as Associate Professor in the ISID team (Information and Decision Systems Engineering) of the CEDRIC laboratory (Centre d'Etude et De Recherche en Informatique et Communications (EA4629)).

I have carried out my research in several projects (WebContent, GeOnto, DataLift, GioQoso, Huma, SAFECARE) in collaboration with academic and industrial partners.

1.1. RESEARCH PROJECTS

Project	Date	Funding	Consortium	Role
SAFECARE: SAFEguard of Critical heAlth infrastructure	2018-2021	EU H2020 (CIP-01-2016-2017)	21 industrial and academic partners including Airbus Cybersecurity SAS, Milestone System AS, AP-HM, Instituto Superior de Engenharia do Porto	*Person in charge of the proposal and task manger*: coordinate the drafting of the project proposal and participate in the implementation of the cyber and physical incident propagation module.
Huma: Humans at the center of massive data analysis for safety	2017-2019	Collaborative projects (FUI)	9 partners, including Airbus Defence & Space CyberSecurity, Intrinsec, Oberthur Technologies, INRIA Nancy	*Member:* participate in the implementation of a trace analysis solution based on Semantic Web technologies.
QUALHIS: Construction and interrogation of large historical prosopographic databases using a quality approach.	2017	CNRS Mastodons	LAMOP – Paris 1 University, ISID CNAM Paris, TECHNE – Poitiers University	*Member:* participate in the design of an ontology-based solution to build and interrogate prosopographic databases.
GioQoso: Quality Management of Open Musical Scores	2016-2018	CNRS Mastodons	CeSR Tours, CNAM Paris, IReMus Paris, IRISA Lannion	*Member:* design and implementation of a solution based on ontologies and inference engines.
DataLift Project: from published raw data to interconnected semantic data	2010-2014	National Research Projects ANR Contint	IGN, INSEE, EXMO, Edelweiss, LIRMM, Eurecom, Atos, Mondeca et FING	*Member:* responsible of implementing the open and linked data portal data.ign.fr
GeOnto: Constitution, alignment, comparison and exploitation of heterogeneous geographical ontologies	2008-2011	National Research Projects ANR MDCO	LRI-IASI, IGN-COGIT, IRIT-IC3, LIUPPA-DESI	*Member:* design and implementation of an alignment solution for refinement and enrichment of heterogeneous geographic ontologies

WebContent: The Semantic Web Framework	2006-2009	National Research Projects ANR RNTL	Fifteen industrial and academic partners including GEMO, MOSTRARE, LIP6, CEA-List, INRA and EADS.	*Member:* implementation of a heterogeneous data alignment solution

1.2. RESEARCH SUMMARY

My work is in the field of the Semantic Web and open data. I am interested in applications of the Semantic Web in which the content of information sources is described using ontologies. More precisely, all of my work done and in progress is grouped around the following research themes:

- ✓ Ontology Partitioning
- ✓ Ontology Alignment
- ✓ Mappings Refinement
- ✓ Ontology Enrichment and Restructuring
- ✓ Data Interlinking in the Web of Data
- ✓ Web of Data Quality
- ✓ Fuzzy time interval representation and reasoning
- ✓ Application of Semantic Web technologies in the fields of security, building and musicology

1.2.1. RESEARCH CARRIED OUT DURING THE MASTER AND PH.D. THESIS

My master and thesis work focuses on the semantic interoperability of heterogeneous information sources and more particularly the alignment of ontologies. They were carried out within the context of the ANR GeOnto project.

1.2.1.1. ONTOLOGY ALIGNMENT AND PARTITIONING

The ontology alignment (identifying mappings or matching concepts) is particularly important in integration systems since it allows the joint consideration of resources described by different ontologies.

This research topic has raised many challenges. Current alignment methods are generally based on measures calculating the similarity of pairs of concepts from the two ontologies. These measures are mostly based on the lexical characteristics of the concept labels and/or the structural characteristics of the ontologies, which involves comparing each concept description from one ontology with the descriptions of all concepts from the other ontology. With the emergence of high-performance and highly expressive representation languages, large ontologies have been developed in complex fields (e.g. medicine, agronomy) and include several tens of thousands of concepts. A possible solution to align these large ontologies is to try to limit the size of the concept sets at the input of the alignment algorithms, by partitioning the two ontologies to be aligned into several blocks, so that only reasonably sized blocks need to be processed.

In the context of my research work, I have proposed two methods that are mainly inspired by co-clustering techniques, which consist in exploiting, in addition to the information expressed by the relationships between concepts within the same ontology, those that correspond to the relationships that can exist between the concepts of the two ontologies. The fact that concepts from both ontologies can have exactly the same label and can be linked by an equivalence relationship is an example of a relationship that is easy to calculate even on large ontologies, and which I have taken advantage of in my proposal.

I performed experiments on ontologies in the geographical domain, provided by COGIT (IGN) in the context of the ANR GeOnto project. These ontologies are well known to the researchers of the team and of limited size, which allows them to be directly aligned with the alignment tool TaxoMap without the need to partition them. This allowed me to analyze the semantic relevance of the generated blocks and to use the results of the direct alignments (without partitions) as a reference for the results obtained after partitioning. Further experiments were then done on a pair of large ontologies.

I also tested the two methods on two large ontologies, AGROVOC and NALT, which are composed respectively of 28439 and 42326 concepts and are used as reference ontologies in the challenge OAEI'08 (Ontology Alignment Evaluation Initiative 2008 Campaign).

The results of this work have been published in [5] [46] [47] [60]

1.2.1.2. MAPPING REFINEMENT

The tests carried out on the taxonomies made available by the IGN's COGIT, a partner in the ANR GeOnto project, showed that the alignment tool TaxoMap (the alignment module) provided very good results (92.3% accuracy), but that these could still be improved. A study of the improvement treatments desired by the experts showed that these were often specific to the aligned ontologies. The objective is not to make TaxoMap a tool solely dedicated to the alignment of topographic taxonomies (the quality of the results would not be guaranteed at all when aligning other ontologies). Therefore, I have proposed an environment allowing experts in a given field to specify and perform refinement treatments on alignments obtained previously. Initially, this environment was used to improve the quality of an alignment provided by TaxoMap (the mapping refinement module). Subsequently, it was used for other treatments based on the results of an alignment between ontologies, such as ontology merging, restructuring or enrichment treatments.

An important feature of the refinement module is to allow a declarative specification of treatments based on the results of a particular alignment, concerning particular ontologies, but using a predefined and generic vocabulary. The processes that can be specified depend on the characteristics of the ontologies concerned and, on the task, to be performed (first the alignment refinement and then the merging, restructuring or enrichment of ontologies). These processes are thus associated with independent specification modules, one for each task, each with its own vocabulary. The module is extensible and a priori applicable to any processing based on alignment results.

I have implemented refinement patterns that I later tested on geographic ontologies, provided by COGIT (IGN) in the context of the ANR GeOnto project.

The results of this work have been published in [43] [44] [45] [46] [58] [59]

1.2.1.3. ONTOLOGY ENRICHMENT

Ontologies and ontology alignment tools are essential components of the Semantic Web since they allow the integration of dispersed sources in a distributed environment. The relationships established between concepts from two distinct ontologies can be used to enrich one of the two ontologies, called target ontology, with the concepts of the other, called source ontology. The ontology enrichment task is composed of two phases: the identification of the relevant concept to be introduced and its placement, with the right relationships, within the target ontology.

The idea I have proposed is to perpetuate some of the identified relationships by aligning the concepts of a source ontology with those of a target ontology, and to introduce some of the concepts from the source as new concepts from the target. The choice of links and concepts of interest for the enrichment of a particular ontology should be made under the responsibility of an expert in the field of the ontology under consideration, but our goal is to provide him/her with an environment capable of assisting him/her in this task. In particular, I have proposed to help the expert sort through the different sources that could be used for enrichment, rejecting those whose themes do not correspond to those of his particular ontology and, if the themes of the two ontologies are compatible, helping him to eliminate mappings that are not very relevant for enrichment.

To do so, I proposed an environment designed to specify enrichment treatments (depending on the enrichment sources) based on mappings produced by an alignment tool. The treatments considered, such as assessing the context proximity of two ontologies or selecting relevant mappings for ontology enrichment, are specifiable for particular ontologies using a set of generic and predefined base primitives.

In my experiments, the target ontology provided by COGIT (IGN) has been enriched with the concepts of (1) an ontology of the same domain and size as the ontology to be enriched, (2) a small ontology from a generalist source (RAMEAU) and (3) a very large generalist ontology (YAGO).

The results of this work were published in [56] [57]

1.2.2. RESEARCH CARRIED OUT DURING POSTDOC

Linked Data is a recent model of distributed data architecture on the Web. Many content providers have already made data available online in Linked Data formats. The interoperability of different datasets published in this way on the Web must be based, on the one hand, on the use of interconnectable schemas through the use of ontologies, and on the other hand, on the creation of effective links between the different resources described by these sets.

One of the objectives of the ANR DataLift project is to help content providers choose the right ontologies to structure their data and to help them interconnect their set with already published sets. A first specificity of this project compared to the numerous proposals being developed in this field is to experiment with the publication of structured reference data provided by traditional content producers (here INSEE and IGN).

I worked on this project on geographic data matching techniques. My task was to propose a generic method of interconnection, applicable firstly to data published on the DataLift platform and secondly to data published on the Web (e.g. DBpedia and GeoNames). I reused the SILK interconnection tool by integrating a new method that mapped data by taking into account their geographical positions (calculated on an ellipsoid). I was also the main actor in the opening of the IGN open data web portal (http://data.ign.fr).

The results of this work were published in [42]

1.2.3. RESEARCH CARRIED OUT AS ASSOCIATE PROFESSOR

My recent research work is in the context of the Semantic Web, where the widespread use of Semantic Web technologies, such as RDF, SPARQL and OWL, has led to the publication of billions of pieces of data on the Web as graphs, called RDF knowledge graphs. In my research, I was interested in the quality of knowledge graphs and their enrichment with contextual identity links and fuzzy temporal data. I am also interested in the application of semantic web technologies in the field of building security and musicology.

1.2.3.1. ENRICHMENT OF GRAPHS WITH DOMAIN-SPECIFIC IDENTITY LINKS

This work was carried out as part of Abdelfettah Feliachi's PhD thesis that I co-supervised with Nathalie Abadie and Bénédicte Bucher (IGN France).

A central issue related to the identification of identity links is the choice of the different parameters used to link entities. For example, in some areas, the use of basic distance measures may not be sufficient to establish links between entities. We have verified this in the context of geographic data that are present in many knowledge graphs published on the Web. In this domain, when the geographic characteristics of features are captured ambiguously, basic spatial mapping algorithms based on distance calculations can produce erroneous links.

To address this issue, we have suggested formalizing and acquiring knowledge about spatial references, namely their planimetric accuracy, geometric modeling, level of detail and imprecision of the spatial features they represent. We then proposed an interconnection approach that dynamically adapts the way spatial references are compared, based on this knowledge.

The results of this work were published in [24] [29] [40] [41] [53] [55]

1.2.3.2. ENRICHMENT OF GRAPHS WITH CONTEXTUAL IDENTITY LINKS

This work was carried out as part of Pierre-Henri Paris' PhD thesis that I co-supervised with Samira Cherfi (Cnam).

Many interconnections between knowledge graphs, published in the Web of Data, use identity links, which assume that the linked entities must be identical in all possible and imaginable contexts. However, since identity depends on context, data describing one entity could be transferred (propagated), erroneously, to another entity via an identity link. Thus, we proposed an approach, based on learning algorithms, to find, in a semi-automatic way, a set of properties, for a given identity context, that can be propagated between contextually identical entities.

The results of this work have been published in [6] [7] [12] [50]

1.2.3.3. ENRICHMENT OF GRAPHS WITH FUZZY TEMPORAL DATA

This work was carried out within the framework of the VIVA project and the PhD theses of Fatma Ghorbel (co-supervised with Elisabeth Métais (Cnam), Nebrasse Ellouze, and Faïez Gargouri (University of Sfax, Tunisia) and Noura Herradi (co-supervised with Elisabeth Métais (Cnam).

The problem we have addressed in these two theses concerns the enrichment of knowledge graphs with imprecise temporal data (for example, late 1970s). This requires the use of fuzzy logic. Using this logic, we proposed an ontology-based approach, allowing representation and reasoning on precise and/or imprecise temporal data. Both quantitative (i.e., time intervals and points) and qualitative (i.e., relationships between time intervals, relationships between a time interval and a point in time, etc.) temporal data were taken into account.

The results of this work were published in [1] [2] [4] [15] [16] [18] [22] [23]

1.2.3.4. KNOWLEDGE GRAPHS QUALITY

This work was carried out as part of Subhi Issa's PhD thesis that I co-supervised with Samira Cherfi (Cnam).

In the academic and industrial world, decision making depends on the quality of data. In this thesis, we proposed to evaluate the quality of knowledge graphs with respect to two important dimensions: completeness and conciseness. For completeness, we proposed a data mining approach to derive a reference pattern (i.e. a set of properties) that will be used in the computation of this dimension. We implemented a prototype, called LOD-CM, to illustrate, for a given knowledge graph, the process of deriving a concept map according to user requirements. Concerning conciseness, its calculation is based on the identification of equivalent predicates. To identify these predicates, we proposed an approach based, in addition to statistical analysis, on a thorough semantic analysis of the data and on learning algorithms.

The results of this work have been published in [14] [19] [25] [51]

1.2.3.5. AP PPLICATION OF SEMANTIC WEB TECHNOLOGIES IN THE FIELD OF SECURITY

This work is being carried out as part of the European H2020 SAFECARE project.

The SAFECAE project addresses the issue of combining cyber and physical threats in the context of health infrastructures which are among the most critical and vulnerable infrastructures. Its main objective is to provide solutions that enhance physical and cyber security in a transparent and cost-effective manner. These solutions will be based on new technologies and innovative approaches to improve threat prevention, detection, incident response and impact mitigation.

Our ISID team at the Cnam CEDRIC laboratory is responsible for the development of the SAFECARE core module, which is responsible for establishing interconnections between the physical and cyber world. This module is based on an impact propagation model that formalizes the relationships between physical and cyber resources, in order to simulate the cascading propagation of effects between these different resources. On the basis of this model, it will be possible to anticipate the potential impacts

of cyber and physical incidents, not only on building infrastructures (power, cooling, etc.), but also on IT systems and medical devices.

We have produced a first version of this module using Semantic Web technologies. A preliminary ontology, which formalizes the concepts of the physical and cyber security domains and their relationships, has been implemented in the OWL 2 language. We then carried out a first simulation of the propagation of impacts on a quasi-real scenario. To do so, we used a reasoner to deduce the propagation of impacts on cyber and physical resources.

1.2.3.6. APPLICATION OF SEMANTIC WEB TECHNOLOGIES IN THE BUILDING DOMAIN

This work is in progress as part of Thamer Mecharnia's thesis that I am co-supervising with Lydia Chibout (CSTB) and Nathalie Pernelle (Université Paris-Saclay).

I am currently collaborating with the Scientific and Technical Centre for Building (CSTB) and the Laboratory for Computer Science at Paris-Saclay University, on the capitalization of knowledge, researchers and experts, in the field of building, in order to perpetuate their knowledge and allow its sharing and reuse in other fields and for other uses. This knowledge must keep its meaning and context to allow an efficient capitalization. This initiative follows on from the work begun in an internal project at the CSTB which aims to offer a support (tool) for identifying asbestos-containing materials, based on a heterogeneous corpus. The objective is to help the operator in the preparation of his identification program and to orient it.

We proposed a first method for predicting the presence of asbestos, based on temporal data, describing the probability of asbestos presence in marketed products. To achieve our goal, we created an asbestos ontology that was populated using data from external resources. This information was then used to calculate the probability of asbestos in the components of a given building.

The first results of this work were published in [10] [52]

1.2.3.7. APPLICATION OF SEMANTIC WEB TECHNOLOGIES IN THE MUSICOLOGY DOMAIN

This work was carried out as part of the CNRS GioQoso project in collaboration with the Vertigo team at CEDRIC, BnF, CESR Tours, iReMus Paris, and the IRISA laboratory in Rennes.

The objective of the GioQoso project is to address issues related to quality in early musical scores. These scores are precious heritage assets, intensively studied by musicological researchers. My role was to explore what the use of Semantic Web technologies could bring to the approaches of quality measurement of musical scores.

As a first step, we proposed a quality management approach based on a formal modeling, in the form of an ontology, of musicologists' expertise in music notation. We then used this model to express contextual rules that can be evaluated, either a priori, to avoid the production of erroneous scores, or a posteriori, to evaluate quality indicators concerning a score or a corpus of scores

In a second step, we proposed an approach that extends the coding of scores with semantic annotations. This approach is based on an ontology of music notation, designed to integrate semantic musical elements, extracted from the score coding or produced by a knowledge extraction process. A mechanism for extracting RDF facts and evaluating semantic rules has been tested on a concrete example, based on the identification of dissonances in a Renaissance counterpoint.

The results of this work have been published in [28] [31]

1.3. SUPERVISIONS

1.3.1. PHD IN PROGRESS

- **Mehdi Zrhal (Supervision rate: 33%)**
  Title: Mediation model for the discovery and reuse of structured geographic content
  *S*upervision: Co-supervised with Bénédicte Bucher and Marie-Dominique Van Damme (IGN France)
  Period: Started in October 2019

- **Nassira Achich (Supervision rate: 25%)**
  Title: Dealing with imperfections in user-entered data - Application to software for Alzheimer's patients
  Supervision: Co-supervised with Elisabeth Métais (Cnam Paris) and Fatma Ghorbel and Faïz Gargouri (Sfax University, Tunisia)
  Period: Started in September 2019

- **Thamer Mecharnia (Supervision rate: 33%)**
  Title: Construction and evolution of a contextual ontology for a multi-purpose building knowledge base
  Supervision: Co-supervised with Nathalie Pernelle (Paris-Saclay University) and Lydia Chibout (CSTB)
  Period: Started in November 2018

1.3.2. DEFENDED PHD

- **Paris Pierre-Henri (Supervision rate: 50%)**
  Title: Identity in RDF knowledge graphs: Propagation of properties between contextually identical entities
  Supervision: Co-supervised with Samira Cherfi (Cnam Paris)
  Period: Started in October 2016 and defended in 17 June 2020

- **Subhi Issa (Supervision rate: 50%)**
  Title: Web data quality: completeness and conciseness
  Supervision: Co-supervised with Samira Cherfi (Cnam Paris)
  Period: Started in October 2015 and defended in 13 December 2019

- **Fatma Ghorbel (Supervision rate: 20%)**
  Title: Intelligent ontology-based graphical dialog for a memory prosthesis
  Supervision: Co-supervised with Elisabeth Métais (Cnam Paris) and Nebrasse Ellouze and Faïz Gargouri (Sfax University, Tunisia)
  Period: Started in November 2015 and defended in 10 July 2018

- **Noura Herradi (Supervision rate: 50%)**
  Title: Multilingual, multicultural and temporal semantic representation of interpersonal relations applied to a memory prosthesis
  Supervision: Co-supervised with Elisabeth Métais (Cnam Paris)
  Period: Started in February 2014 and defended in 20 December 2018

- **Abdelfettah Feliachi (Supervision rate: 50%)**
  Title: Interlinking and visualizing georeferenced resources of the Web of data with geographic reference data
  Supervision: Co-supervised with Nathalie Abadie and Bénédicte Bucher (IGN France)
  Period: Started in September 2013 and defended in 27 October 2017

1.4. SEMINARS AND TALKS

- A Multilingual Semantic Similarity-Based Approach for Question-Answering Systems, In the International Conference on Knowledge Science, Engineering and Management (KSEM 2019), Aug. 28, 2019, Athens, Greece.
- An approach for measuring RDF data completeness, In the French Conference on Advanced Databases (BDA 2015), Sept. 29, 2015, France.
- GeomRDF: A Fine-Grained Structured Representation of Geometry in the Web, In GeoLD'14 workshop, Sept. 1, 2014, Leipzig, Germany.
- Data Linking. In EGC'14 tutorial presentations, Jan. 28, 2014, Rennes, France.
- Linked geographic Data and INSPIRE: the Datalift experience. In OGC - Open Data Working Group, Mar. 06, 2013, Saint Mandé, France.
- A practical introduction to geographical linked data: lift your data. In INSPIRE'12 workshop presentations, Infrastructure for Spatial Information in Europe, Jun. 24, 2012, Istanbul, Turkey.
- TaxoMap Framework: ontology alignment and mappings refinement. May. 3, 2012, within a visit to the CRIM (Centre de Recherche Informatique de Montréal), Montréal, Canada.
- Enhancing the semantic interoperability: applicability and utility of the ontology alignment. Apr. 2, 2012, within a visit to the LIPN Lab, University of Paris 13, France.
- Alignment of heterogeneous geographic ontologies. May. 23, 2011, within a visit to the INSERM U936 du Lab, Rennes, France.
- Enhancing the semantic interoperability trough the ontology alignment: applicability and reuse. Apr. 22, 2011, within a visit to the LIRMM Lab, University of Montpellier II, France.
- TaxoMap Framework applied to the alignment of geographic ontologies in the GeOnto project. In OntoGeo workshop presentations at SAGEO'10 conference, Nov. 18, 2010, Toulouse, France.
- Pattern-based Mapping Refinement. In EKAW'10 conference presentations, Oct. 12, 2010, Lisbon, Portugal.
- WikiLinkr: Semi-automatic linking for Semantic MediaWiki. In the SSSC - IEEE 2010 Summer School on Semantic Computing, Jul. 31, 2010, Berkeley, California.
- A Framework for Mapping Refinement Specification, in AISB'10 symposium presentations, Mar. 31, 2010, De Montfort University, Leicester, UK.
- Scraping for research. In the SSSW - The Seventh Summer School on Ontological Engineering and the Semantic Web, Jul. 11, 2009, Cercedilla, near Madrid, Spain.
- Alignment-based Partitioning of Large-scale Ontologies. In EGC'10 conference presentations, Jan. 29, 2010, Strasbourg, France.

1.5. REVIEWER FOR

- SWJ'20 - Semantic Web Journal
- ICONIP'19 - International Conference on Neural Information Processing
- ER'19, 20 - International Conference on Conceptual Modeling
- KSEM'19 - International Conference on Knowledge Science, Engineering and Management
- EGC'17, 18, 19, 20 - Extraction et la Gestion des Connaissances.
- DKE'17, 18 – Data & Knowledge Engineering
- AICCSA'17 - 14th ACS/IEEE International Conference on Computer Systems and Applications
- WEBIST'17, 18, 19 - International Conference on Web Information Systems and Technologies
- SoWeDo'16 - Workshop in IC'16 conference (Knowledge engineering)
- TSI'16 - Technique et Science Informatiques french journal
- RNTI-OPEN-DATA'15 - Special issue of the RNTI journal
- IJCAT'14 - International Journal of Computer Applications in Technology

- FGCS Journal'14 - Future Generation Computer Systems
- INS'13 - Informatics and Computer Science Intelligent Systems Applications, International Journal, Elsevier
- SIGMOD Record'13 - The ACM Special Interest Group on Management of Data
- ISI'13 - Special issue of the ISI French journal (Information Systems Engineering) - Evaluation of Information Systems
- AKDM'10 - Advances in Knowledge Discovery and Management, Studies in Computational Intelligence, Springer
- COSI'10 - Colloque sur l'Optimisation et les Systèmes d'Information.
- EGC-M'10 - Conférence Maghrébine sur l'Extraction et la Gestion des Connaissances.
- OntoGeo'10 - Workshop, in SAGEO'10 conference (Spatial Analysis and GEOmatics)

1.6. ORGANIZATION COMMITTEE OF:

- Co-Director of the special issue "The Web of Data: Publication, Linking and Capitalization", RSTI-ISI (Hermès- Lavoisier)
- "Quality of Linked Data (QLOD)" Workshop, in EGC'16, EGC'17 conferences
- "Linking in the Web of Data" Tutorial, in EGC'14 conference
- "A practical introduction to geographical linked data: lift your data" Workshop, in INSPIRE 2012, Infrastructure for Spatial Information in Europe, June 24, 2012, Istanbul, Turkey.
- "Ontologies Géographiques (OntoGeo)" Workshop, in SAGEO'10 conference (Spatial Analysis and GEOmatics). Toulouse, November 2010.

1.7. TOOLS & SOFTWARES

GeomRDF: a tool that transforms geospatial data from traditional GIS to RDF. It is based on a vocabulary that reuses and extends GeoSPARQL and NeoGeo so that geometries can be defined in any CRS, and represented as both structured and GeoSPARQL-compliant geometries.

TaxoMap: an automatic ontology alignment tool.

TaxoMap Web Service: a web service implementation of the TaxoMap tool.

The TaxoMap Framework: an environment designed to help an expert in the field to specify treatments based on product mappings, in order to refine them or to merge, restructure or enrich ontologies.

AlignViz: a tool for visualizing alignments between two ontologies.

TaxoPart: a tool for partitioning ontologies.

These tools can be downloaded from my website: https://cedric.cnam.fr/~hamdif/ - software

The TaxoMap Web Service is available at
http://taxomap.lri.fr:8090/axis2/services/TaxoMapService?wsdl
The web service client is deployed at: http://taxomap.lri.fr/

1.8. PUBLICATIONS

International Journals

[1] Ghorbel, F., Hamdi, F. and Métais, E., "Dealing with Precise and Imprecise Temporal Data in Crisp Ontology", International Journal of Information Technology and Web Engineering (IJITWE), 15(2), pp.30-49, 2020.

[2] Ghorbel, F., Hamdi, F., Métais, E., Ellouze, N. and Gargouri, F., "Ontology-Based Representation and Reasoning about Precise and Imprecise Temporal Data: A Fuzzy-Based View", Data & Knowledge Engineering, vol. 124 (November 2019), pp. 26, 2019.

[3] Métais, E., Ghorbel, F., Herradi, N., Hamdi, F., Lammari, N., Nakache, D., Ellouze, N., Gargouri, F. and Soukane, A., "Memory Prosthesis", Non-pharmacological Therapies in Dementia, vol. 3(2), pp. 177-180, 2015.

[4] Metais, F. Ghorbel, F. Hamdi, N. Ellouze, N. Herradi, A. Soukane, "Representing Imprecise Time Intervals in OWL 2", Enterprise Modelling and Information Systems Architectures - International Journal of Conceptual Modeling, vol. 13, pp. 120-132, 2018.

Book Chapters

[5] Hamdi, F., Safar, B., Reynaud, R. and Zargayouna, H., "Alignment-based Partitioning of Large-scale Ontologies", Chapter of the book Advances in Knowledge Discovery and Management, Studies in Computational Intelligence, Guillet, F.; Ritschard, G.; Zighed, D.A.; Briand, H. (Eds.), Springer, 2010, p. 251-269.

International Conferences/Workshops

[6] Paris, P.H., Hamdi, F. and Cherfi, S.S., "OntoSemStats: An Ontology to Express the Use of Semantics in RDF-based Knowledge Graphs". In Proceedings of Web Engineering - 20th International Conference, ICWE 2020, Helsinki, Finland, June 9-12, 2020.

[7] Paris, P.H., Hamdi, F. and Cherfi, S.S., "A Study About the Use of OWL 2 Semantics in RDF-Based Knowledge Graphs". In the proceedings of the ESWC 2020 Satellite Events, 2020.

[8] Ghorbel, F., Wali, W., Hamdi, F., Métais, E. and Gargouri, B., "Visualizing Readable Instance Graphs of Ontology with Memo Graph", ICONIP 2019, December 2019, pp.12 pages, Sydney, Australia.

[9] Ghorbel, F., Wali, W., Métais, E., Hamdi, F. and Gargouri, B., "A Smart Search Functionality for a Memory Prosthesis: A Semantic Similarity-Based Approach", International Conference on Digital Health Technologies (ICDHT 2019), December 2019, Hammamet, Tunisie.

[10] Mecharnia, T., Pernelle, N., Khelifa, L.C. and Hamdi, F., "An Approach Toward a Prediction of the Presence of Asbestos in Buildings Based on Incomplete Temporal Descriptions of Marketed Products", K-CAP'19, November 2019, pp.4, Marina del Rey, California, USA.

[11] Achich, N., Ghorbel, F., Hamdi, F., Metais, E. and Gargouri, F., "A Typology of Temporal Data Imperfection", The 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (KEOD 2017), September 2019, pp.6, Vienna, Austria.

[12] Paris, P.H., Hamdi, F. and Cherfi, S.S., "Interlinking RDF-based datasets: A structure-based approach", 23rd International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2019), September 2019, pp.11, Budapest, Hungary.

[13] Wali, W., Ghorbel, F., Gragouri, B., Hamdi, F. and Metais, E., "A Multilingual Semantic Similarity-Based Approach for Question-Answering Systems", The 12th International Conference on Knowledge Science, Engineering and Management (KSEM 2019), August 2019, pp.8, Athens, Greece.

[14] Issa, S., Hamdi, F. and Cherfi, S.S., "Enhancing the Conciseness of Linked Data by Discovering Synonym Predicates", The 12th International Conference on Knowledge Science, Engineering and Management (KSEM 2019), August 2019, pp.12, Athens, Greece.

[15] Achich, N., Ghorbel, F., Hamdi, F., Metais, E. and Gargouri, F., "Representing and Reasoning about Precise and Imprecise Time Points and Intervals in Semantic Web: Dealing with Dates and Time Clocks", 30th International Conference on Databases and Expert Systems Applications (DEXA 2019), August 2019, pp.10, Linz, Autriche.

[16] Ghorbel, F., Hamdi, F. and Métais, E., "Temporal Relations between Imprecise Time Intervals: Representation and Reasoning", 24th international conference on conceptual structures (ICCS 2019), July 2019, pp.14, Marburg, Germany.

[17] Ghorbel, F., Hamdi, F. and Métais, E., "Estimating the Believability of Uncertain Data Inputs in Applications for Alzheimer's Disease Patients", 24th International Conference on Applications of Natural Language to Information Systems (NLDB 2019), June 2019, pp.12, Series LNCS, Salford, United Kingdom.

[18] Ghorbel, F., Hamdi, F. and Métais, E., "Ontology-Based Representation and Reasoning about Precise and Imprecise Time Intervals", FUZZ-IEEE 2019 International Conference on Fuzzy Systems, June 2019, pp.8, New Orleans, USA.

[19] Issa, S., Paris, P.H., Hamdi, F. and Cherfi, S.S., "Revealing the Conceptual Schema of RDF Datasets", 31st International Conference on Advanced Information Systems Engineering (CAiSE), June 2019, pp.1-15, Italy.

[20] Ghorbel, F., Métais, E., Hamdi, F., Ellouze, N. and Gargouri, F., "A Memory Training for Alzheimer Patients", SETIT 2018, December 2018, pp.10 pages, Genoa & Hammamet, Italy & Tunisia.

[21] Ghorbel, F., Métais, E., Hamdi, F., Ellouze, N. and Gargouri, F., "An Incremental Extraction and Visualization of Ontology Instance Summaries with Memo Graph", SETIT 2018 Science of Electronics, technologies of Information and Telecommunication, December 2018, pp.10 pages, Genoa & Hammamet, Italy & Tunisia.

[22] Ghorbel, F., Hamdi, F., Métais, E., "A Crisp-Based Approach for Representing and Reasoning on Imprecise Time Intervals in OWL 2", ISDA 2018 (Intelligent Systems design and Applications), December 2018, pp.12 pages, Vellore, India.

[23] Ghorbel, F., Hamdi, F., Métais, E., Ellouze, N. and Gargouri, F., "A Fuzzy-Based Approach for Representing and Reasoning on Imprecise Time Intervals in Fuzzy-OWL 2 Ontology", NLDB 2018, June 2018, pp.12, Paris, France.

[24] Feliachi, A., Abadie, N. and Hamdi, F., "Assessing the Positional Planimetric Accuracy of DBpedia Georeferenced Resources", In Proceedings of ER Workshops 2017: 227-237.

[25] Issa, S., Paris, P.H. and Hamdi, F., "Assessing the Completeness Evolution of DBpedia: A Case Study", In Proceedings of ER Workshops 2017: 238-247.

[26] Navarro, J., Legrand, V., Lagraa, S., François, J., Lahmadi, A., De Santis, G., Festor, O., Lammari, N., Hamdi, F., Deruyver, A. and Goux, Q., "HuMa: A Multi-layer Framework for Threat Analysis in a Heterogeneous Log Environment", In Proceedings of FPS 2017: 144-159.

[27] Herradi, N., Hamdi, F. and Métais, E., "A Semantic Representation of Time Intervals in OWL 2", In Proceedings of KEOD 2017.

[28] Cherfi, S.S., Guillotel, C., Hamdi, F., Rigaux, P. and Travers, N., "Ontology-Based Annotation of Music Scores", In Proceedings of K-CAP 2017: 10:1-10:4.

[29] Feliachi, A., Abadie, N. and Hamdi, F., "An Adaptive Approach for Interlinking Georeferenced Data", In Proceedings of K-CAP 2017: 12:1-12:8

[30] Ghorbel, F., Hamdi, F., Ellouze, N., Métais, E. and Gargouri, F., "Visualizing Large-scale Linked Data with Memo Graph",, In Proceedings of the 21st International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2017), September 2017, pp.10, Marseille, France.

[31] Cherfi, S., Hamdi, F., Rigaux, P., Thion, V. and Travers, N., "Formalizing Quality Rules on Music Notation - an Ontology-based Approach", In Proceedings of the International Conference on Technologies for Music Notation and Representation (TENOR 2017), May 2017, pp.1--7, Coruna, Spain.

[32] Ghorbel, F., Métais, E., Ellouze, N., Hamdi, F. and Gargouri, F., "Towards Accessibility Guidelines of Interaction and User Interface Design for Alzheimer's Disease Patients", in the Tenth International Conference on Advances in Computer-Human Interactions (ACHI 2017), March 2017, pp.6, Nice, France.

[33] Ghorbel, F., Ellouze, N., Métais, E., Hamdi, and F., Gargouri, F., "MEMO_Calendring: a Smart Reminder for Alzheimer's Disease Patients", In Proceedings of the International Conference on Smart, Monitored and Controlled Cities (SM2C 2017), February 2017, pp.8, Sfax, Tunisie.

[34] Ghorbel, F., Ellouze, N., Métais, E., Hamdi, F., Gargouri, F. and Herradi, N., "MEMO GRAPH: An Ontology Visualization Tool for Everyone", In Proceedings of the 20th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2016), September 2016, pp.10, Series Procedia Computer Science, York, UK.

[35] Ghorbel, F., Ellouze, N., Métais, E., Gargouri, F., Hamdi, F. and Herradi, N., "Designing and Evaluating Interfaces for the CAPTAIN MEMO Memory Prosthesis", In Proceedings of ACHI 2016 (The Ninth International Conference on Advances in Computer-Human Interactions), April 2016, pp.16, Venice, Italy.

[36] Herradi, N., Hamdi, F., Métais, E. and Soukane, A., "PersonLink: A Multilingual and Multicultural Ontology Representing Family Relationships", In Proceedings of the 7th International Conference on Knowledge Engineering and Ontology Development (KEOD 2015), November 2015, pp.8, Lisbon, Portugal.

[37] Herradi, N., Hamdi, F., Ghorbel, F., Métais, E., Ellouze, N. and Soukane, A., "Dealing with Family Relationships in Linked Data", In the IEEE International Workshop on Computational Intelligence for Multimedia Understanding (IWICM 2015), October 2015, pp.5, Prague, Czech Republic.

[38] Herradi, N., Hamdi, F., Métais, E., Ghorbel, F. and Soukane, A., "PersonLink: An Ontology Representing Family relationships for the CAPTAIN MEMO Memory Prosthesis", In Proceedings of ER 2015 (Workshop AHA), October 2015, Vol. 9382 , pp.3-13, Series Lecture Notes in Computer Science, Stockholm, Sweden.

[39] Mrabet, Y., Hamdi, F. and Métais, E., "Towards Targeted Open Data Interlinking", In the IEEE International Workshop on Computational Intelligence for Multimedia Understanding (IWCIM 2014), november 2014, Paris, France.

[40] Hamdi, F., Abadie, N., Bucher, B. and Feliachi, A., "GeomRDF: A Fine-Grained Structured Representation of Geometry in the Web", In the first International Workshop on Geospatial Linked Data (GeoLD 2014), 1 september 2014, Leipzig, Germany.

[41] Feliachi, A., Abadie, N., Hamdi, F. and Atemezing, G. A., "Interlinking and visualizing linked open data with geospatial reference data", In Proceedings of the 8th International Workshop on Ontology Matching co-located with the 12th International Semantic Web Conference (ISWC 2013), Sydney, Australia, October 21, 2013.

[42] Scharffe, F., Euzenat, J., Villata, S., Troncy, R., Hamdi, F., Bucher, B., Cotton, F., Gandon, F., Fan, Z., Bihanic, L., Kepeklian, G., Atemezing, G., Vandenbussche, P. and Vatant, B., "Enabling linked-data publication with the Datalift Platform", In Proceedings of the AAAI workshop on semantic cities, Toronto (ONT CA), 2012.

[43] Hamdi, F., Safar, B., Niraula, N. and Reynaud, R., "TaxoMap alignment and refinement modules: Results for OAEI 2010", In Proceedings of the Ontology Alignment Evaluation Initiative (OAEI) 2010 Campaign - ISWC Ontology Matching Workshop, Shanghai International Convention Center, Shanghai, China, November 7, 2010.

[44] Hamdi, F., Reynaud, C. and Safar, B., "Pattern-based Mapping Refinement", In Proceedings of the 17th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2010), 11th October-15th October 2010, Lisbon, Portugal.

[45] Hamdi, F., Reynaud, C. and Safar, B., "A Framework for Mapping Refinement Specification", In Proceedings of the International Symposium on Matching and Meaning, Michael Chan and Fiona McNeill (Eds.), at the AISB 2010 convention, 29 March - 1 April 2010, De Montfort University, Leicester, UK.

[46] Hamdi, F., Safar, B., Niraula, N. and Reynaud, R., "TaxoMap in the OAEI 2009 alignment contest", In Proceedings of the Ontology Alignment Evaluation Initiative (OAEI) 2009 Campaign - ISWC Ontology Matching Workshop, Westfields Conference Center near Washington, DC, 2009.

[47] Hamdi, F., Zargayouna, H., Safar, B. and Reynaud, R., "TaxoMap in the OAEI 2008 alignment contest", In Proceedings of the Ontology Alignment Evaluation Initiative (OAEI) 2008 Campaign - ISWC Ontology Matching Workshop, Karlsruhe, 2008.

National Journals

[48] Hamdi, F. and Cherfi, S.S., "Une approche pour évaluer la complétude de données RDF", Dans Ingénierie des Systèmes d'Information, 21(3), pp.31-52, 2016.

[49] Hamdi, F. and Saïs, F., "Éditorial", Dans Ingénierie des Systèmes d'Information, 21(5-6), pp.7-10, 2016.

National Conferences/Workshops

[50] Paris, P.H., Hamdi, F. and Cherfi, S.S., "Etat des lieux de l'utilisation de OWL 2 : Analyse et proposition pour capturer les utilisations de la sémantique OWL 2 dans les graphes de connaissances RDF", Dans les actes de la conférence Extraction et Gestion des Connaissances (EGC), RNTI-E-36 : 145–156, 2020.

[51] Issa, S., Paris, P.H., Hamdi, F. and Cherfi, S.S., "Revealing the Conceptual Schemas of RDF Datasets - Extended Abstract", Dans les actes de INFORSID 2020.

[52] Mecharnia, T., Pernelle, N., Khelifa, L.C. and Hamdi, F., "Approche de prédiction de présence d'amiante dans les bâtiments basée sur l'exploitation des descriptions temporelles incomplètes de produits commercialisés", Journées francophones d'Ingénierie des Connaissances (IC 2019), July 2019, pp.14, Toulouse, France.

[53] Feliachi, A., Abadie, N. and Hamdi, F., "Que représentent les références spatiales des données du Web ? un vocabulaire pour la représentation de la sémantique des XY", In Proceedings of Atelier Qualité des Données du Web (QLOD'16) Joint à EGC 2016, January 2016, pp.7-12, France.

[54] Hamdi, F. and Cherfi, S. C., "An approach for measuring RDF data completeness", Bases de Données Avancées (BDA 2015), September 2015, pp.1-10, France.

[55] Feliachi, A., Abadie, N. and Hamdi, F., "Intégration et visualisation de données liées thématiques sur un référentiel géographique", 14ème Conférence Francophone sur l'Extraction et la Gestion des Connaissances (EGC2014). Rennes, 28 janvier - 31 janvier, 2014.

[56] Hamdi, F., Safar, B. and Reynaud, R., "Extraction de sous-parties ciblées d'une ontologie généraliste pour enrichir une ontologie particulière", Dans les actes de la 12ème Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances (EGC2012), Revue des Technologies de l'Information - RNTI. Bordeaux, 31 janvier - 3 février, 2012.

[57] Hamdi, F., Safar, B. and Reynaud, R., "Utiliser des résultats d'alignement pour enrichir une ontologie", Dans les actes de la 11ème Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances (EGC2011), Revue des Technologies de l'Information - RNTI, pages 407-412. Brest, 25-28 janvier, 2011.

[58] Hamdi, F., Reynaud, C. and Safar, B., "TaxoMap Framework appliqué à l'alignement d'ontologies géographiques dans le projet GéOnto", Dans les actes de l'atelier OntoGeo associé à SAGEO'2010, pp. 51-53, Toulouse, 18 novembre 2010.

[59] Hamdi, F., Reynaud, C. and Safar, B., "L'approche TaxoMap Framework et son application au raffinement de mappings", Dans les actes du congrès Francophone Reconnaissance des Formes et Intelligence Artificielle, RFIA 2010, Caen, 19-22 janvier 2010.

[60] Hamdi, F., Safar, B., Zargayouna, H. and Reynaud, R., "Partitionnement d'ontologies pour le passage à l'échelle des techniques d'alignement", Dans les actes 9èmes Journées Francophones sur l'Extraction et Gestion des Connaissances (EGC 2009), Revue des Technologies de l'Information - RNTI, pages 409-420. Strasbourg, 27-30 Janvier, 2009. (Best Application Paper Award)

Theses

[61] Hamdi, F., "Enhancing the semantic interoperability: applicability and utility of the ontology alignment", Ph.D in Computer Science, University of Paris-Sud 11. December 2011.

Technical Reports

[62] Cherfi, S.S., Lammari, L., Hamdi, F. and Atigui, F., "Specification of the impact propagation and DS models", Deliverable 6.6, SAFECARE H2020 Project, December 2019.

[63] Hamdi, F., Morel, C., Reynaud, R. and Safar, B., "Comparison of ontology alignment tools", Deliverable D5, DataBridges ICT Labs activity, November 2011.

[64] Hamdi, F., Reynaud, C. and Safar, B., "Ontology Alignment", Final report, ANR MDCO GeOnto Project, February 2011.