Université Grenoble Alpes

Habilitation à Diriger des Recherches

Spécialité informatique et mathématiques appliquées

Measures for knowledge

with applications to Ontology Matching and Data Interlinking

Jérôme David

Date de soutenance : 15 mai 2023

Composition du jury

Présidente :	Marie-Christine Rousset, Professeur, Université Grenoble Alpes
Rapporteurs :	Marianne Huchard, Professeur, Université de Montpellier
	Axel-Cyrille Ngonga Ngomo, Professeur, Universität Paderborn
	York Sure-Vetter, Professeur, Karlsruher Institut für Technologie
Examinatrice :	Nathalie Pernelle, Professeur, Université Sorbonne Paris Nord

Chapter 1

Introduction

1.1 Context

Most of the data exposed on the web is interpretable by humans but not directly by computers. Let us imagine that you are visiting a page about the painting 'The Weeping Woman'. You will probably see a picture that you will interpret as a picture of a painting that represents a woman. By reading the text around, you will also learn that this is a painting of Pablo Picasso.

The semantic web is an extension of the web enabling people to express knowledge in a way that machines can reason with it. If the knowledge about 'The Weeping Woman' were expressed with semantic web technologies, an application could have exploited it. For instance, an application could have deduced that 'The Weeping Woman' is a portrait painting by using previous knowledge and other facts expressed on other pages such as 'a portrait painting is a painting depicting (at least) a human' and 'a woman is a human'. Since its proposal in 1998, the semantic web has gained a lot of popularity and it is now a reality. For instance, if you write the query 'subject of The Weeping Woman' in some well-known search engine, it will retrieve the entity 'Dora Maar'. And if you ask for the 'model of Pablo Picasso', it will find also 'Dora Maar' among other entities. These are possible because this search engine makes use of a large knowledge base that describes classes (Painting, Artists, etc.), properties ('has painted', 'subject', etc.), and instances ('Dora Maar', 'The Weeping Woman', etc.).

In the semantic web, knowledge is expressed through ontologies. **Ontologies** are structures that define formally classes, properties and their relationships. Classes and properties are usually organized into subsumption hierarchies. For instance, an ontology can define the class Painter as a subclass of Artist and it can also set the domain of the 'has painted' property to Painter. OWL and RDFS are the two main languages for defining ontologies. The descriptions of instances are called data and can be formalized thanks to RDF.

There may be several ontologies about the same field. If one wants to make them interoperable, the relations between ontologies have to be discovered. This task is called **ontology matching** and it results in a set of correspondences asserting the relations between classes and properties is called an alignment.

For example, consider two organizations, a bookstore and a national library, that use ontologies represented on Figures 1.1 and 1.2 using a (simplified) Manchester syntax of OWL2 (Horridge and Patel-Schneider, 2012). Even if the two organisations describe books, they do not have the same goals and then they do not use the same ontologies.

```
Class: Auteur
Class: Editeur
Class: Livre

ObjectProperty: auteur
Domain: Livre
Range: Auteur

ObjectProperty: editeur
Domain: Livre
Range: Editeur

ObjectProperty: sujet
Domain: Livre
Range: owl:Thing

DataProperty: nom
Domain: Auteur or Editeur
Range: xsd:string

DataProperty: prix
Domain: Livre
Range: xsd:decimal

DataProperty: titre
Domain: Livre
Range: xsd:string
```

Figure 1.1: Bookstore ontology

```
Class: Artwork
Class: Manifestation
Class: Person

Class: Autobiography
   SubClassOf: Biography

Class: Biography
   EquivalentTo:
      Book and (focus some Person)

Class: Book
   SubClassOf: Artwork

Class: Painter
   EquivalentTo:
      inverse (contributor) some Painting

Class: Painting
   SubClassOf: Artwork

Class: PortaitPainting
  EquivalentTo:
     Painting and (depicts only Person)

ObjectProperty: contributor
  Domain: Artwork
  Range: Person

ObjectProperty: depicts
  SubPropertyOf: focus
  Domain: Painting

ObjectProperty: focus
  Domain: Artwork
   Range: owl:Thing

ObjectProperty: hasManifestation
   Domain: Artwork
   Range: Manifestation

DataProperty: birthdate
   Domain: Person

DataProperty: birthplace
   Domain: Person

DataProperty: name
   Domain: Person

DataProperty: title
   Domain: Artwork
```

Figure 1.2: National library ontology

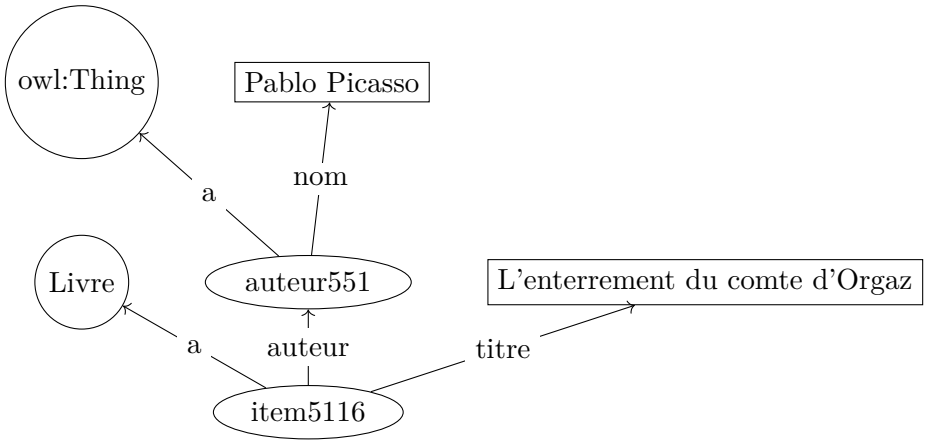


Figure 1.3: Bookstore instances

To get a better presentation of his catalogue, the book seller may want to exploit the ontology provided by the national library. To that extent, the seller ontology and the national library ontology have to be aligned. An alignment between these ontologies would contain:

- 'sujet' is more specific than 'focus'
- 'auteur' is more specific than 'contributor'
- 'Livre' is a subclass of 'Book'

Thanks to this alignment, the bookstore can refine his topics and better structure his catalogue For instance, using the alignment and the national library ontology, it could be inferred that a 'Livre' having a 'Person' as one of its 'sujet' is a biography. When ontologies are aligned, knowledge expressed on both sides can interoperate and be consolidated. However, it may happen that ontologies are not aligned or that the relations between them are too vague to be useful.

Similarly, instances may be represented by various organisations in different way. It is thus necessary to identify the same instances between different knowledge bases and publish theses links. This task is called **Data interlinking** and can be thought of as complementary to ontology matching. For instance, Figures 1.3 and 1.4 show data described with respectively the bookstore and the national library ontologies. Let us imagine that the bookseller wants to retrieve books (instance of class 'Livre') written by a painter. Since the bookstore does not have any class 'Painter', using only the alignment between the bookstore and the national library ontologies is not sufficient to select such books. If one can identify that the instance 'auteur551' of the bookstore is the same as the instance 'Picasso' of the national library, then it become possible to infer than 'item5116' is a book written by a painter. This entailment does not only use ontologies but also alignments and links between the bookstore and the national library.

It first shows, on a single example, the intricacies of exploiting heterogeneous knowledge and the complementary role played by ontology, data, alignments and links. Then our

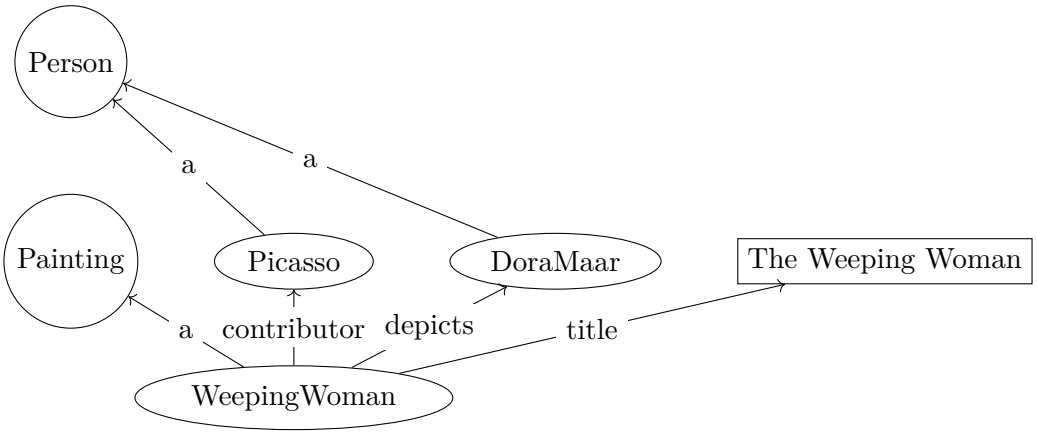


Figure 1.4: National library instances

objective is not to reduce the diversity of knowledge but to understand heterogeneity in order to benefit from it.

1.2 Problems and approaches

Ontology matching and **data interlinking** are cornerstones of semantic web technologies. They enable knowledge to be shared, consolidated and reused across organizations. Many methods for ontology matching and data interlinking have been proposed. But no solution solves theses problems in all scenarii. In fact, they are not easy tasks because ontologies may vary both in term of granularity, semantics and natural language used in annotations.

Ontology matching and data interlinking rely on tools able to automatically discover correspondences between concepts (classes and properties) and links between instances. Their general underlying principle is to compare entities (classes, properties, instances) and to decide the relations that hold between them, if any. Relations between classes and properties can be equivalence, subsumption, overlapping or disjointness. Between instances, we will concentrate only on the equality relation expressed through the OWL constructor owl:sameAs.

In my work since my PhD (2007), I addressed the problems of comparing and measuring distances between representations of knowledge. One can distinguish between tree levels of representation: (1) ontologies, (2) alignments and (3) instances.

Ontologies. A first problem is to measure the proximity between ontologies. Since there are different scenarii that need to establish proximity between ontologies, a single measure will not cover all the requirements. For instance, if one wants to complete her ontology, she will search for other ontologies that share some concepts with it but that have not necessarily the same axioms. Meanwhile, if the objective is to exchange data between applications, the more axioms are shared by the ontologies, the easier the exchange will be. One of the difficulties resides in providing several operational and relevant definitions of proximities.

Two approaches for measuring proximities could be considered: relying only on the

content of ontologies (ontology space) or using alignments as support of the similarity (alignment space). For the search use case, measures defined on the ontology space are suitable because they do not require an alignment to be available. In some other applications, the similarity has to be driven by the alignments. For instance, if one wants to quantify the coverage of query mediation between ontologies, then the coverage will depend on the alignments used by the application. In this scenario, a proximity measure should reflect the overlaps between ontologies given by the alignment. Designing alignment-space similarity measures requires to deal with the whole network of alignments. In particular, the relation between two concepts is not necessarily directly asserted but can be deduced considering several paths of correspondences and relations within ontologies.

Alignments. A second challenge is the evaluation of the quality of alignments between ontologies. The classical way to assess the quality of an alignment consists in comparing it with a reference alignment. This is usually performed by representing alignments as simple sets of correspondences and by comparing these sets using the precision and recall measures. This basic approach neither considers the semantics of ontologies and alignments nor the proximity between correspondences. In fact, two alignments that are syntactically different, i.e. do not share any correspondence, could be semantically equivalent. Furthermore, if we take the semantics into account we have to relax the all-or-nothing nature of logical entailment: It is not because two alignments are not semantically equivalent that they are not close. This is thus a difficult task to redefine precision and recall measures in order to both consider semantics and proximity. The approach that I followed has been to study the properties that precision and recall have to satisfy in order to respectively measure the correctness and completeness of alignments.

When no reference alignment is available, assessing the quality of an alignment is even more difficult. There are several options for approximating this through satisfying specific properties: consistency, conservativity, locality (Solimando, Jiménez-Ruiz, and Guerrini, 2014; Solimando, Jiménez-Ruiz, and Guerrini, 2017). As mentioned previously, correctness and completeness are two meaningful indicators of quality (Meilicke, 2011). Correctness can be guessed using consistency or satisfiability checking. Completeness is more difficult to measure when the complete alignment is not known. In both cases, several semantics, RDFS and OWL, have to be addressed. Our approach assesses the correctness and completeness of alignment by measuring how axioms or relations between concepts of one ontology can be translated into the second one modulo alignments.

Instances. The last part of my work tackles the challenge of identifying resources from different RDF datasets that represent the same entity. In fact the same entity can be described differently in RDF datasets because they may use different ontologies. Many works have investigated the use of similarities for comparing instances from different datasets (Nikolov, Ferrara, et al., 2011; Nentwig, Hartung, Ngonga Ngomo, et al., 2017a). Identifying resources can also take advantage of keys and alignments. A key is a set of properties that characterize uniquely all instances of a given class. For instance, if a key holds for a data set and if this key can be translated to a second dataset using an alignment between the ontologies of the two data sets, then this key can be used as a rule for identifying the same instances across these data sets. However, keys are rarely asserted at ontology level and we have also shown (Atencia, Chein, et al., 2014) that other definitions of a key than those of the owl:hasKey construct can be useful for data interlinking. We have first addressed the problem of inducing keys from data. Discovering keys from data has been solved for tables from the relational database model, but in the case of RDF we have to deal with the non functionality of properties.

Link discovery can be addressed by combining both keys and alignments. However, alignments are not necessarily available or there may be no pair of aligned keys. My approach is to relax the key condition to be valid only on the intersection of the data sets. We have followed this idea to develop the notion of link key as a generalization of a pair of keys related by an alignment. In particular, we have tackled the discovery of such link key patterns from data. The discovery method only requires as input the two datasets, i.e. without alignments. However the possible number of sets of pairs of properties to consider is exponential. Furthermore, we have to be able to select only relevant ones. As a consequence, we have focused on developing techniques to reduce the search space and also measures to assess a priori the quality of extracted link key patterns.

1.3 Contributions

In order to deal with heterogeneity, I have studied and contributed to techniques and measures for comparing knowledge structures on the semantic web. We are interested in three kinds of knowledge structures: ontologies, alignments and instances. The main originality of my work is to always focus on taking advantage of the semantics of knowledge structures.

Ontologies. For comparing ontologies, we have investigated two families of measures: ontology space and alignment space measures. The first family of measures only relies on the content of the ontologies whilst alignment space measures take advantage of existing alignments between ontologies.

We have provided and compared several ontology space similarity measures (David and Euzenat, 2008b). Our proposed measures mainly differ on the deepness of the information they take into account. The lightest measures only rely on text annotation of concepts, while the deepest ones compare the triples shared by ontologies.

We have also developed totally new alignment space measures and we have studied them experimentally (David, Euzenat, and Sváb-Zamazal, 2010). More precisely, we have proposed path-based measures which only consider the existence of paths of alignments between ontologies in a network and coverage based-measure which evaluates the proportion of entities that are covered by paths of alignments.

Alignments. We have contributed to the evaluation of alignments by investigating two approaches to evaluate the semantic quality of alignments with a reference one or without a reference. For the first approach, we have studied how to combine both semantics and proximity of correspondences. Our contribution has been to provide a framework that generalizes both semantic (Euzenat, 2007; David and Euzenat, 2008b) and relaxed (Ehrig and Euzenat, 2005) alignment evaluation measures.

A second approach consists in evaluating the quality of alignment without a reference one. We have investigated both syntactic and semantic translations by generalizing agreement and disagreement measures (d'Aquin, 2009). Since such alignment quality measures are usable in any non-controlled evaluation contexts, they can be very useful in practice.

Instances. We have proposed symbolic approaches to perform data-interlinking by developing an algorithm to extract pseudo-keys from RDF data (Atencia, David, and Scharffe, 2012) and also by introducing the notion of a link key (Atencia, David, and Euzenat, 2014a). We have addressed the discovery of RDF pseudo-keys by extending a method for extracting functional dependencies from relational databases. We have provided definitions of keys and pseudo-keys for RDF and developed an efficient algorithm for extracting them. Keys can not only be used for interlinking data but also for cleaning

data. In fact, it happens that counter-examples of a pseudo-key are often duplicates within the data set.

Regarding link keys, we have designed an algorithm that exploits both the notion of candidate link key and indexing techniques. We have developed measures for assessing the quality of extracted candidates and have shown that this approach gives good results in practice. We have formalized the extraction problem with formal concept analysis (FCA) and extended it to interdependent link key extraction using relational concept analysis (RCA) (Atencia, David, Euzenat, et al., 2020). Finally, we have theoretically compared the different semantics of keys and link keys (Atencia, David, and Euzenat, 2021). This work demonstrates that link keys are more general than a pair of keys whose properties are related by alignments.

Moreover, all these contributions have been implemented in open source software and libraries, some of which have been widely used.

1.4 Editorial choice and manuscript structure

The goal of this document is to put my work in perspective. Rather than a historical perspective, I chose to organise it as a structured synthesis, highlighting the relations and the lines of forces of my work. I have done my best to minimize the formalisms used in this manuscript and have tried to use examples to illustrate important notions.

As stated previously, I have addressed three kinds of knowledge structures: ontologies, alignments and instances. The structure of this document follows this segmentation.

Chapter 2 is dedicated to ontology measures. Section 2.3 gives an overview of ontology space measures that take as input only the ontologies without using alignments. In Section 2.4, we present alignment space measures, and in particular measures based on the existence of correspondences between ontologies and measures that quantify the coverage of entities through alignments.

Chapter 3 concerns alignment evaluation measures. We make the distinction between extrinsic and intrinsic quality measures. Extrinsic evaluation uses external information such as a reference alignment, while intrinsic evaluation relies solely on the alignment and content of ontologies. In Section 3.2 about extrinsic evaluation, we present and discuss our generalisation of semantic and relaxed precision and recall. For intrinsic measures, we propose, in Section 3.3, inclusion and exclusion measures generalizing agreement and disagreement.

Chapter 4 deals with the data-interlinking problem. Section 4.2 presents an algorithm for extracting pseudo-keys, i.e. keys allowing few exceptions, from RDF datasets. It shows that it is useful for data-interlinking but also for cleaning data, i.e. identifying duplicates. In Section 4.3, we extend the notion of keys to link keys. In particular, we give an overview of algorithms for extracting link key candidates and evaluation measures that we have developed.

We conclude by providing perspectives of this work. So far, we have worked on static knowledge and we have provided methods and tools for reconciliating it. Knowledge and data may and have to evolve. In this context, we will study how a network of aligned ontologies and linked data have to adapt to changes.

1.5 Origin of materials

This manuscript is mainly based on works presented in the following papers. Since I have chosen to present a synthetic view of my work, these papers may provide more precise and technical details.

Chapter 2 is based on two papers, the first presents measures in the ontology space, and the second those in the alignment space:

- Jérôme David and Jérôme Euzenat (2008a). "Comparison between ontology distances (preliminary results)". In: *Proc. 7^{th} international semantic web conference (ISWC), Karlsruhe (DE)*. vol. 5318. Lecture notes in computer science, pp. 245–260. URL: https://exmo.inria.fr/files/publications/david2008a.pdf

- Jérôme David, Jérôme Euzenat, and Ondrej Sváb-Zamazal (2010). "Ontology similarity in the alignment space". en. In: *Proc. 9^{th} international semantic web conference (ISWC), Shanghai (CN)*, pp. 129–144. URL: https://exmo.inria.fr/files/publications/david2010b.pdf

Chapter 3 presents work that not been published yet but that is an extension of the following paper:

- Jérôme David and Jérôme Euzenat (2008b). "On fixing semantic alignment evaluation measures". en. In: *Proc. 3^{rd} ISWC workshop on ontology matching (OM), Karlsruhe (DE)*. ed. by Pavel Shvaiko et al., pp. 25–36. URL: https://exmo.inria.fr/files/publications/david2008b.pdf

Chapter 4 summarizes results on data-interlinking published in several conference and journals. The first paper addresses pseudo key discovery, and the second the link key discovery. The third paper presents the interdependent link key extraction through Relational Concept Analysis (RCA). The forth one study composition of link keys. Finally the last one formalizes the different semantics of link keys and compare them with key.

- Manuel Atencia, Jérôme David, and François Scharffe (2012). "Keys and pseudo-keys detection for web datasets cleansing and interlinking". en. In: *Proc. 18th international conference on knowledge engineering and knowledge management (EKAW), Galway (IE)*, pp. 144–153. URL: https://exmo.inria.fr/files/publications/atencia2012b.pdf

- Manuel Atencia, Jérôme David, and Jérôme Euzenat (2014a). "Data interlinking through robust linkkey extraction". en. In: *Proc. 21^{st} european conference on artificial intelligence (ECAI), Praha (CZ)*. ed. by Torsten Schaub et al. Amsterdam (NL): IOS press, pp. 15–20. URL: https://exmo.inria.fr/files/publications/atencia2014b.pdf

- Manuel Atencia, Jérôme David, Jérôme Euzenat, et al. (2020). "Link key candidate extraction with relational concept analysis". en. In: *Discrete applied mathematics* 273, pp. 2–20. URL: https://moex.inria.fr/files/papers/atencia2019z.pdf

- Manuel Atencia, Jérôme David, and Jérôme Euzenat (2019). "Several link keys are better than one, or extracting disjunctions of link key candidates". en. In:

Proc. 10th ACM international conference on knowledge capture (K-Cap), Marina del Rey (CA US), pp. 61–68. URL: https://moex.inria.fr/files/papers/atencia2019c.pdf

- Manuel Atencia, Jérôme David, and Jérôme Euzenat (2021). "On the relation between keys and link keys for data interlinking". en. In: *Semantic web journal* 12.4, pp. 547–567. URL: https://content.iospress.com/articles/semantic-web/sw200414

Contents

Chapter 2

Ontology Measures

We are interested in providing measures of proximity between ontologies. To do this, we need to define functions that take two ontologies as input and return a number quantifying the similarity or dissimilarity of the two ontologies. At first glance, this seems simple, but they are many dimensions and constraints to consider.

There could be many different scenarii where there is a need for evaluating proximity between ontologies. For instance, one could think to the knowledge engineer verifying if the ontology she is developing will easily interoperate with other ontologies on the same domain. In this case, an ontology measure taking advantage of the structure and the semantics is suitable. Another use case, is the librarian who indexes books with thesauri. He could be interested to enrich the indexation by using other theasauri on the same domain. In this scenario, the structure of related thesauri does not matter but they have to share similar terms in their annotations. We could also imagine a company that does not want to disclose the core of its ontology. However, in order to exchange with customers and suppliers, the company exposes a minimal interface of its ontology with alignments between some shared domains ontologies. Somebody willing to know if his ontology is close or not of those of the company can only rely on alignments without having access to the content of the targeted ontology.

As we can see, the target application plays a primary role in the design of a measure because each application has his own constraints and uses different level of representation. We addressed the problem of measuring proximity between ontologies following this idea that there is not a single perfect measure but several ones that ground on different aspects of ontologies such as the textual content, the structure or the semantics. We investigated separately measures in the ontology space relying only the ontology content and those in the alignment space that take advantage of alignments between ontologies. In both cases, we have identified several ways to compare ontologies and designed different measures.

2.1 Related work

There are a lot of different works related to similarities and ontologies. Many similarities have been developed assessing the relatedness of concepts within an ontology (Pirrò, 2019) or for matching ontologies (Euzenat and Shvaiko, 2013). These are different approaches as the former is defined on a single ontology, while the latter applies to several ontologies.

The first family of ontology measures, usually named ”semantic similarities“, takes advantage of the ontology in order to quantify how concepts are similar. A measure can be structural in the sense that it depends directly on the length of the paths connecting

the concepts of the ontology. Measures based on information theory, which consists in quantifying the amount of information that two concepts have in common, have also been proposed. Some of them use corpus-based concept probabilities (Resnik, 1995), others estimate the amount of information from the subsumption hierarchy (Seco et al., 2004; Pirrò and Euzenat, 2010).

The second family of ontology measures, such as those of (Mädche and Staab, 2002), (Hu et al., 2006) and (Vrandečić and Sure, 2007), widely used in ontology matching is in reality concerned with concept distances across ontologies. (Mädche and Staab, 2002) introduced a concept similarity based on terminological and structural aspects of ontologies. This very precise proposal combines an edit distance on strings and a structural distance on hierarchies (the cotopic distance). This ontology similarity strongly relies on the terminological similarity. OLA (Euzenat and Valtchev, 2004) uses a concept similarity for ontology matching. This measure takes advantage of most of the ontological aspects (labels, structure, extension) and selects the maximum similarity. It is thus a good candidate for ontology similarity. The framework presented in (Ehrig, Haase, et al., 2005) provides a similarity combining string similarity, concept similarity – considered as sets – and similarity across usage traces.

There is also a quite elaborate framework in (Hu et al., 2006). This paper is mostly dedicated to the comparison of concepts but can be extended to ontologies. First, concepts are expanded so that each concept is expressed as a disjunction of compound but conjunctive primitive concepts. This works as long as no cycle occurs in the ontology. Then primitive concepts are considered as dimensions in a vector space and each concept is represented in this space. The weights used in this vector space are computed with TF·IDF. The distance between two concepts is the smallest cosine distance between vectors associated with disjuncts describing concepts.

Finally, (Vrandečić and Sure, 2007) more directly considered metrics evaluating ontology quality. This is nevertheless one step towards semantic measures since they introduce normal forms for ontologies which could be used for developing syntactically neutral measures.

These works generally rely on elaborate distance or similarity measures between concepts. Their extension to distances between ontologies is never discussed, although there are many ways to do so.

The works cited above consider only measures defined on the ontology content. However, one could assume that ontologies are already aligned, and take advantage of these networks of aligned ontologies in order to define measures. A first step toward alignment space measure as been proposed by (d'Aquin, 2009). This paper investigated ontology agreement which is used as a measure for choosing compatible ontologies. It can be seen as another kind of distance or similarity between ontologies. However, the way agreement/disagreement is computed in the cited paper is still mainly based on ontology content and alignments are only used for identifying equivalent entities which are not immediately comparable.

In Section 3.3, we specifically adapt and generalize these two measures in the context of ontology alignment evaluation.

2.2 Terminology

Ontologies

In the semantic web, the word "ontology" refers to different structures according to the context. It can be simple SKOS thesaurus (Miles and Bechhofer, 2009), RDFS vocabularies (Brickley and Guha, 2014) or expressive OWL ontologies (Group, 2012).

To simplify, we use here a simple but generic definition of an ontology. It can be of course refined according to the context, but this definition is sufficient to understand the idea of the proposed measures.

Definition 1 (Simple ontology)**.** *An ontology is defined as a tuple* $O = \langle E, \mathcal{T}, \mathcal{L} \rangle$ *where*

- E *is the set of entity names (IRIs) defined in the ontology. It encompasses classes, properties, and individuals.*
- $\mathcal{T}$ *is the set of logical axioms (A-Box and T-Box) of the ontology.*
- $\mathcal{L}$ *is the set of annotation axioms.*

Considering the representation model $\mathcal{T}$ can be either a set of RDF triples or a set of OWL axioms. Annotations $\mathcal{L}$ can be human-readable text annotations such as `rdfs:label`, `rdfs:comment` and `skos:prefLabel`, but also other kind of annotations such as `rdfs:seeAlso` or `owl:deprecated`. To simplify, we will consider that $\mathcal{L}$ contains only text annotations. The semantics of ontologies can be characterised by entailment, denoted $\models$, which enables to define the closure of an ontology $Cn(O) = \{\delta; O \models \delta\}$.

This way of defining an ontology makes it possible to highlight different levels on which similarity measures can be based.

An ontology space is characterized by a set of ontologies.

Alignments

Alignments express correspondences between entities belonging to different ontologies. Here we will only use a simplified version of alignments; a more complete definition and discussion can be found in (Euzenat and Shvaiko, 2007).

Definition 2 (Simple alignment)**.** *Given two ontologies* o *and* o'*, with their associated set of entities* E *and* E' *, an* alignment *is a set of correspondences* $\langle e, e', r \rangle$*, such as:*

- $e \in E$ *and* $e' \in E'$ *are entities issued from the ontologies;*
- $r \in \mathbf{R}$ *is the relation that holds between* e *and* e'

A correspondence $\langle e, e', r \rangle$ asserts that the relation r holds between the ontology entities e and e'. In most of the cases, the set of relations $\mathbf{R}$ is equivalence ($=$) and subsumption ($\sqsubseteq, \sqsupseteq$), but one could also relies on algebras of alignment relations (Inants, 2016).

We call alignment space a set of ontologies related by alignments between them. It can be seen as a multi graph where vertices are ontologies and edges are alignments. In alignment spaces, pairs of ontologies are not necessarily directly aligned. However, if there is a path between two ontologies in the space, then an alignment can be computed by composing alignments from the path.

Dissimilarity, similarity and distance

Ontology measures can be dissimilarities, similarities or distances, and they differ according to their properties. In this context of this work, all measures are functions defined on a given ontology-space or alignment-space. These functions take as input two ontologies and return a non-negative real number.

A dissimilarity is a symmetric function which is minimal, i.e. equals to 0, when applied to the same ontology. A distance is a dissimilarity which also satisfies the triangular inequality and is equal to 0 only if the ontologies are the same.

The similarity is a dual operation of dissimilarity : it is as large as the ontologies are similar. In our work, we mainly consider normalized similarities that are maximal and equal to 1 when applied to same ontology.

2.3 Ontology-space measures

When only ontologies are available without alignment between them, a similarity measure can be based on the ontology content. In the following, we call a set of ontologies without alignment, an ontology space. This section presents a summary of the ontology space measures that we have proposed and compared.

Because there are different requirements for evaluating proximity between ontologies, measures can rely on different levels of the ontology content. A first dimension is the type of content on which the measure is defined:

- a **text-based measure** relies on the annotations axioms.
- a **structural measure** will consider logical axioms as a graph.
- a **semantic measure** take the semantics of the ontology language into account.

This first category is not disjoint from the structural and semantic categories and a measure can be mixed with both approaches.

A second dimension is the granularity of the measure: we make the distinction between **global measures** that consider ontology as a whole single structure from **concept-based measures** that compare each pair of concepts separately and then aggregate the results into one single value.

We saw that most of the work dealing with measures between ontologies comes from ontology matching and these measures are in fact defined between concepts from different ontologies. Such concept-based measures consist in computing a value between concepts and then aggregate all theses values. As a second category we consider global measures that do not compare concepts separately but the whole set of annotations and/or axioms as a single structure.

Finally, we could differentiate intrinsic measures only relying on the ontological content from extrinsic measures that make use of external sources of information or knowledge. External resources can be lexical databases such as WordNet (Miller, 1995) or word-embedding models, but also aligned ontologies (Locoro et al., 2014). Except for the last one, for which we have proposed and studied several measures (see Section 2.4), we did not specifically studied such extrinsic measures.

In (David and Euzenat, 2008a), we proposed and compared several global and concept-based measures. These measures are discussed in the following sections and a classification

Measures	Type of content			Granularity	
	Text-based	*Structural*	*Semantic*	*Global*	*Concept-based*
CosineVM	x			x	
JaccardVM	x			x	
Lexical	x				x
Triple-based	x	x			x

Table 2.1: Classification of proposed measures regarding their type of content they rely on and their granularity.

of them according to the type of content used by measures and their granularity is given Figure 2.1.

Global text-based measures

A similarity or dissimilarity can be computed by comparing the sets of labels appearing in both ontologies and using a measure such as the Hamming distance, i.e., the complement to 1 of the ratio of common terms over the whole set of terms used by any of the ontologies. This distance would certainly run faster than any serious matching algorithm but does not tell a lot about the matching process.

However, more elaborate measures based on the vector space model (VSM) have been designed. The approach consists in representing an ontology as a bag of terms extracted from the annotations and making use of classical information retrieval metrics in the vector space model. In particular we have compared a baseline measure based on the number of term in common (Jaccard index on boolean model) with the classical cosine index with TF·IDF weights.

Concept-based measures

Another approach represents an ontology as a set of concepts. These concepts will depend on the techniques used for establishing the distance: they will generally be the classes or properties to be found within the ontologies. In this case, defining a distance between the ontologies often relies on:

- a distance (δ) or similarity (σ) measure between concepts;
- a collection distance (Δ) which uses the distance between concepts for computing a distance between ontologies.

In our work (David and Euzenat, 2008a), we have designed two new concept-based measures and analysed their associations with several collection measures independently. The first one only relies on lexical information from annotations. It represents each entity by a set of string values. Then the similarity is computed from the map that maximizes the Jaro-Winkler similarity between the two sets of string values.

The second measure mixes both textual and structural informations. It relies on a representation of input ontologies as RDF graphs and is defined between RDF nodes. The idea is that the similarity between two nodes depends on how similar the nodes appearing in their respective neighbourhood are. Initially, the similarity of literal nodes is given by a syntactic measure and those of nodes having the same URI is set to one. Then the

similarity between other nodes is iteratively updated using the values computed at the previous iteration.

Once one has a similarity (or dissimilarity) among concepts available, there are different choices for extending measures at the concept level to the ontology level. This is achieved with the help of a collection measure which computes the ontology measure value from the concept measure values.

As collection measures, we have considered the average linkage, the Hausdorff distance and a minimum weight maximum graph matching distance. The problem with the Hausdorff distance and linkage measures, is that its value is only function of the distance between one pair of members of the sets. The average linkage, on the other hand, is function of the distance between all the possible comparisons. None of these are satisfactory. Matching-based dissimilarities (Valtchev, 1999) measure the dissimilarity between two ontologies by taking into account an alignment (matching) between these two ontologies. It can be defined independently of any alignment by using the minimum weight maximum matching. This last measure considers the problem of distance between sets of entities as an assignment problem. In particular, this measure consists in computing a (injective) mapping between two sets of concepts that minimizes the distance (or maximizes the similarity) and then makes an average of individual values of this mapping. The implementation relies on the Hungarian algorithm (Kuhn, 1955).

Evaluation and results

At the time of such work, measures had, to the best of our knowledge, not been evaluated as ontology distances. We had emitted opinion on their relevance only grounded on their mathematical form. It was necessary to enhance this judgement through evaluation. We have evaluated both the speed of distance computation and the accuracy with regard to asserted similarity.

An ideal experimental setting comprises a corpus of ontologies with clear expectations about the distances that should be found between them. We do not have such a corpus annotated with distances values between ontologies. However, the most important thing is to know the proximity order between ontologies.

Finding a relevant corpus was not an easy task. We have chosen the OAEI benchmark suite because it offers a collection of ontologies systematically altered from one particular ontology. From this test set, we have been able to build a reference partial order between the ontologies. The experiments have consisted in comparing the partial order given by measures to the reference one obtained from alterations.

Results showed that the triple-based similarity performed the best. Surprisingly, all global lexical measures were performing better than the concept-based lexical measures. The cosine measure combined with TF weights obtained very good results close to those of the triple-based similarity.

In term of runtime, as expected, global measures were faster than concept-based measures: they were 5 times faster to compute than the lexical concept-based measure and 80 times faster than the triple-based similarity (David and Euzenat, 2008a).

Extensions of this work

We have not pursued directly this work on ontology space distances. However, the lesson that we learned from designing measures and experimental results have been useful for other works. For instance, in the T. Lesnikova's PhD thesis (Lesnikova, 2016), we have

investigated measures in the context of cross-lingual data-interlinking. The challenge was to design fast, scalable and accurate lexical measures for comparing instances whose annotations were expressed in different natural language. Starting for the observations that both triple-based similarity and global measures based on VSM and cosine distance were accurate, we have combined both. Then, an instance is represented as a collection of tokens extracted from annotations coming from the instance itself but also from entities that are closely related in the graph.

2.4 Alignment-space measures

So far we have discussed ontology-space distances defined only on the content of ontologies. However, one could take advantage of the context around ontologies. In this section, we are interested in comparing ontologies using alignments expressing relations between concepts in these ontologies (Euzenat and Shvaiko, 2007). More specifically, a distance or similarity measure is alignment-based if it is computed without relying on the content of ontologies, but only on that of the alignments. So, such a measure can only be applied when alignments are available, we assume that the semantic web contains many ontologies already available and some alignments, sometimes competing, between them. We call alignment space such a structure populated by ontologies related by alignments.

Alignment space measures may seem more remote from the true distance between the ontologies because they do not directly consider their content. However, there are cases where they can be very useful. This is obviously the case when ontologies are not available, e.g. because they are on a closed server, but alignments between these ontologies and others exist. Such unavailable ontologies may be used as a target ontology or as an intermediate ontology (and then alignments may be composed).

This is also the case when the similarity between ontologies has to reflect the ability to transform a statement or a query from one ontology to another, e.g. in semantic peer-to-peer systems or dynamic composition of semantic web services. Since alignment spaces are structured by actual alignments, an alignment space measure is indeed reflecting to some extent the capacity to translate ontology expressions. Such measures would be even more useful if they could be computed quickly with respect to a particular query or formula. On the other hand, distances in an ontology space only provide a measure of closeness, and an alignment or a mediator remains to be produced.

Even if ontologies are available, such measures may be useful as approximations of the "real distance" which are easier to compute than comparing the ontologies: alignment-based measures can quickly provide a hint on what are the most promising options. Indeed, because they already provide the structure to compute the measure, alignments are faster to compare than elaborate comparison of two ontologies as a whole.

Path-based measures

At an upper level an alignment space can be represented as a graph whose vertices are ontologies and edges indicate the existence of an alignment between the connected ontologies. An example of such a graph is given Figure 2.1. A first kind of similarity between two ontologies may be based on paths in this graph.

In fact, the existence of a path guarantees that an alignment between the two ontologies can be computed and that queries could be translated from one ontology to another. A first basic measure considers different values if the path is made of zero, one or several

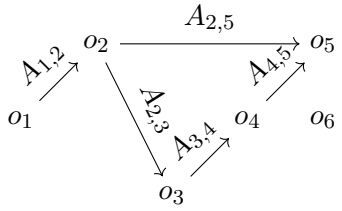


Figure 2.1: An example of a graph at the granularity of alignments given by an alignment space

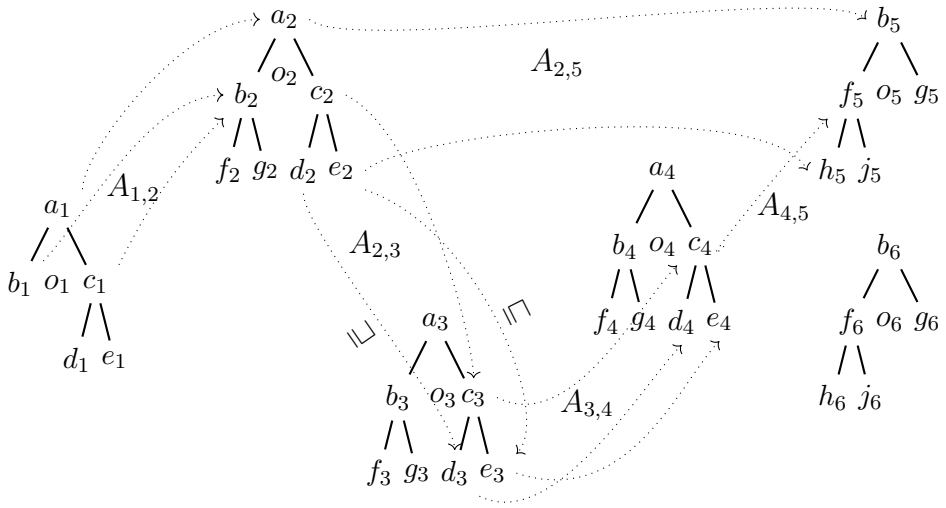


Figure 2.2: A network composed of six ontologies and 5 alignments

alignments. The similarity is maximal when the ontologies are the same and minimal when there is no alignment between them, for instance between o_1 and o_6. The similarity is higher when there is a single direct alignment, e.g. between o_1 and o_2, than when several alignments have to be composed, e.g. between o_1 and o_4.

However, this first similarity is not very precise in the number of transformations that may have to be performed to propagate this information. For example, o_1 is as similar to o_3 as to o_4. So, a natural extension of this measure depends on the shortest path in the alignment space. Indeed, the fewer alignments are applied to a query, the more it is expected that it is an accurate translation (in first approximation). In this case, the similarity between o_1 and o_3 will be higher than those between o_1 and o_4 because the last one requires at least 3 alignments instead of 2.

These path-based measures are very easy to compute but they do not quantify precisely how ontologies are similar. For instance, there could exist a path of alignments between ontologies that leads, through composition, to an empty alignment. And even if alignments are not empty, this measure does not tell how much of an ontology is preserved through the translation.

Coverage-based measures

If we want to go further in measuring the precise proximity for querying applications, it may be useful to consider the ratio of concepts of one ontology which are covered by an alignment. In fact this can be applied to any set of elements, not just an ontology. Hence the coverage can be given with regard to an ontology entity (the ratio is 1 or 0), to a query or to an ontology. It corresponds to the percentage of entities which have an image through a given alignment. Alignments may not be functional nor injective: several concepts can be matched to a single one and vice-versa. There is a second important notion which is the ability for the alignment to preserve the difference between concepts which are deemed different in the source ontology. We have then introduced the alignment distinguishability measure defined as the proportion of matched entities which are kept distinct. Both coverage and distinguishability can be merged into a single measure that is the ratio between the number of matched concepts over the number of source concepts. For instance, on Figure 2.2, the coverage of o_1 through $A_{1,2}$ is 3/5, the distinguishability is 2/3 and the coverage-distinguishability is 2/5.

In the ontology space, there is not necessarily an alignment between two ontologies but a path of alignments. In this case, an alignment can be computed as the composition of the alignments along this path.

Following this idea of coverage, we have proposed two measures. The first one will be computed on the path that maximizes the coverage-distinguishability while the second one is calculated on an alignment obtained by the union of all paths between the two ontologies. For instance, the first measure between o_2 and o_5, that only use the alignment $A_{2,5}$, is equal to 2/7 while the second one, that also exploits the path $A_{2,3}, A_{3,4}, A_{4,5}$, is equal to 3/7. These measures are not symmetric, but they can be easily transformed into similarities.

Experiments and results

These measures in the alignment space have been compared to those proposed in the ontology space. In (David, Euzenat, and Sváb-Zamazal, 2010), we have performed several experiments using the OntoFarm dataset (Šváb et al., 2005).

The most interesting aspect of the results is that coverage-based measures were far more correlated with the measures in the ontology space than to the path-based measures. They were even more correlated to the global measures in the vector-space than the global measures agree with the concept-based measures. This shows that the coverage measures, which do not have access to the content of ontologies, are meaningful with regard to this content.

The two path-based measures were poorly correlated to the others because of the topology of the alignment space: either there exist as short path between a pair of ontologies or they are not connected at all. The graph of alignments in OntoFarm is very connected (91 alignments out of 105 possible ones) so these measures are not discriminant: the ontologies come in few groups depending on how they are connected to the others, most of them being reachable through one alignment.

Moreover, we have also shown experimentally that coverage measure are reasonably robust to errors in the alignments, especially if individual correspondences are missing. This is very encouraging for measures that do not rely on ontology content.

2.5 Conclusions and perspectives

This chapter presented advances made on measuring distances or similarities between ontologies. We distinguished the cases where ontologies are aligned or not. In both cases, we have proposed several measures, analysed their characteristics and evaluated them experimentally. Results have shown that global measures based on vector space are simple yet relevant. Especially these measures are also more correlated to coverage-based measure in the alignment space, than more elaborate concept-based measures.

We have not pursued directly on this topic for the last ten years, but lessons learned from these works have partly guided some of our work, in particular that of Tatiana Lesnikova's thesis (Lesnikova, 2016). With the recent development of ontology embeddings (Chen et al., 2021), our global ontology measures could be redefined in the space of embedding and compared to those we proposed.

More generally, we believe that diversity help to produce more robust knowledge and should be taken into account by ontology matching, knowledge completion, and ontology learning methods. Indeed, there is not one single way to learn knowledge and evaluating methods only in term of accuracy is not sufficient. Diversity in knowledge representation has not been thoroughly studied so far. However, diversity allows knowledge to be more adaptable in new situations. For instance, it has been shown, in different contexts, that groups of agents with diverse abilities have better problem solving skills those with high abilities (Stirling, 2007; Hong and Page, 2004; Noble et al., 2015). In the more general context of knowledge evolution through communication between agents, we considered measuring and controlling knowledge diversity of a population of agents (Bourahla, David, et al., 2022). In particular we considered to measure diversity by taking advantage of distance between ontologies such as those discussed in this chapter.

My perspective is to pursue the study of measuring knowledge diversity, taking advantage of ontology distances and diversity measures developed in the context of phylogenetics (Tucker et al., 2016; Leinster, 2021). There are different point of views of diversity and then several measures have been developed. The simplest one, the α-diversity is the number of species that are observed. Another family of measure also consider the abundance of each species. These measures, such as the Hill numbers, are mostly based on the concept of entropy. A last family of measures takes into account similarity between species. Diversity measures and relations between have been also studied from mathematical point of view (Leinster, 2021).

In a first case, we are interested in measuring the diversity of concepts within a single ontology. The simplest measure of diversity is the number of concepts within the ontology. But, as for phylogenetics, more elaborate measures taking into account the number of instances per concepts and similarity between concepts can be developed. Such measures have applications for verifying and controlling the diversity when a single ontology evolves.

In the second case, one could be interested in evaluating the diversity of a set of agents based on their knowledge (ontology). This can be useful, for instance, in multi-agent simulations such as those performed for cultural knowledge evolution (Bourahla, Atencia, et al., 2021). Such measures have to count how many different ontologies are used by a population of agents, their frequency of usage and also the similarity between them. Controlling diversity in cultural knowledge evolution experiments would help to analyse the impact of the diversity on the communication success and also on the community resilience in the face of disruptive events.

Chapter 3

Quality measures for ontology alignments

Since there is a variety of ontologies, the problem of matching ontologies is difficult to solve in a universal way. Ontologies may vary in terms of formalization in terms of text annotations, natural language used for annotations, granularity, etc. In this context, many ontology matchers have been proposed (Euzenat and Shvaiko, 2013).

To assess their strengths and weaknesses, ontology matchers have to be evaluated. The Ontology Alignment Evaluation Initiative (OAEI)[1] aims at providing a consensus for the evaluation of ontology matchers. Since 2004, OAEI organizes, every year, an evaluation event and publishes the results. This helps the developers of ontology matchers to improve their systems and allows everyone to compare matching strategies on an open and common basis (Achichi, Cheatham, et al., 2016).

The evaluation of ontology alignments often relies on comparing them with a reference alignment. To assess their quality, classical precision and recall measures are usually computed. It has been shown that these measures suffer from several drawbacks since they do consider neither the semantics of aligned ontologies nor the atomic proximity between found correspondences and expected ones. To overcome such limitations, relaxed precision and recall (Ehrig and Euzenat, 2005) and semantic precision and recall (Euzenat, 2007) have been proposed.

Precision and recall-based measures are well suited in controlled environments where reference alignments are available. But in real scenarios, these reference alignments are of course not known. In that context, the quality of ontology alignments can be approximated using different heuristics such as consistency (or satisfiability) (Meilicke, 2011) or conservativity principle (Solimando, Jiménez-Ruiz, and Guerrini, 2017).

In our work, we were interested in the evaluation of alignments without considering the dimensions relative to the tools that have produced them, such as runtime efficiency, or scalability. We also do not consider the particular task for which the alignments will be used such as query answering (Solimando, Jiménez-Ruiz, and Pinkel, 2014). In the following, we make distinction between extrinsic evaluation relying on a reference alignment from intrinsic evaluation that could only take advantage of two ontologies and an alignment between them.

Our work have addressed both kinds of evaluation approaches. On the extrinsic part, our contributions is a generalization of both relaxed and semantic frameworks that allows

[1] http://oaei.ontologymatching.org

to repair some of their individual problems. For intrinsic evaluation, we have adapted and generalized the principle of agreement and disagreement between ontologies (d'Aquin, 2009).

3.1 Related work

Extrinsic evaluation

When we worked on alignment evaluation, the most common way to evaluate ontology alignments was to compare them with a reference and calculate precision and recall measures (Euzenat, 2003). These measures consider alignments as set of correspondences and compare them strictly, using the size of the intersection between both sets, without considering neither proximity between correspondences nor their semantics.

Starting from the fact that it is not because a correspondence is not in the reference that it is fully incorrect, (Ehrig and Euzenat, 2005) has proposed to relax precision and recall measures. Classical precision and recall take as numerator the size of the intersection between the evaluated and the reference alignments. Instead of this, a relaxed measure use an overlap function which quantifies how the alignments are close to each other. The overlap function aggregates proximities between correspondences and can be one of the collection distances discussed in Section 2.3, such as the maximum weight maximum matching measure. The paper proposes three concrete measures: symmetric, correction effort and oriented proximities (one for each precision and recall). The symmetric proximity is function of the distances between entities and the correction effort adopts the approach of an edit distance between correspondences. The oriented proximities are defined differently for precision and recall. They are a first attempt to capture the semantics of alignments in the sense that if an evaluated correspondence is entailed by a reference one (but have the same relation), the proximity will be 1 for precision. However, as in the classical model, these measures are still syntactic measures and do not satisfy any of the properties desired by a semantic model such as a precision equals to 1 when the evaluated alignment can be deduced from the reference one. But, they can be a good basis for new measures which partially consider semantic.

(Euzenat, 2007) introduced semantic precision and recall for evaluating alignments. In particular, the author first introduces ideal measures that are defined on the deductive closures of alignments (named α-consequences) instead of the set of asserted correspondences. However, such sets can be infinite hence the measure could not be calculated. Then the author has proposed the semantic precision to be the proportion of asserted correspondences that can be deduced from the reference alignment and the semantic recall to be the proportion of reference correspondences than can be deduced from the evaluated alignment.

We have shown in (David and Euzenat, 2008b) that theses measures can be artificially increased by introducing redundancy and we have proposed to fix these measures using ideal semantic precision and recall bounded to a set of alignments.

Measures based on alignment repair strategies

The evaluation of alignments without a reference can be supported by the principles of consistency and conservativity discussed in (Jiménez-Ruiz et al., 2011).

(In)consistency in alignments has been introduced by (Stuckenschmidt et al., 2006) in the context of reasoning with alignments. (Meilicke and Stuckenschmidt, 2008; Meilicke,

2011) defined the degree of incoherence as the ratio between the size of a/the minimal set of correspondences that introduce incoherency over the the size of the alignment. The complement to 1 of this degree of incoherence is an upper bound for the precision, if we require the reference alignment to be coherent. Coherency, here means that all the concepts that were satisfiable in each ontology separately are still satisfiable on the merged ontology, i.e. the union of the ontologies and the alignment. From this (Jiménez-Ruiz et al., 2011) has defined the consistency principle which states that the theory resulting of the union of ontologies and their alignment should be consistent and the all the named entities (named classes, data properties and object properties) should be satisfiable.

The conservativity principle states that an alignment should not introduce new relations between concepts of input ontologies. The literature highlights two variations on this principle. (Jiménez-Ruiz et al., 2011) states: given ontologies o and o' and alignment a between o and o', then $o \cup a$ should not introduce new relations between concepts of o. (Solimando, Jiménez-Ruiz, and Guerrini, 2017) gives an even more stronger definition: $o \cup o' \cup a$ should not introduce new relations between concepts of o. This last work also introduces two relaxations of the conservativity principle: subsumption and equivalence conservativity principles.

The general conservativity principle is a strong statement based on the intuition that if an alignment allows to deduce new knowledge then this is probably incorrect. This is, to our opinion, very restrictive because an alignment can be beneficial for both ontologies involved. However, in (Solimando, Jiménez-Ruiz, and Pinkel, 2014), it has been shown, in the context of query rewriting, that it can impact the quality of the results and it has also been integrated as a quality indicator in various OAEI tracks.

Instead of focusing solely on potential errors, it may make sense to quantify the common and divergent knowledge that an alignment brings out. This is the principle of agreement and disagreement measures presented in the following section.

Instance-based evaluation

Another way to overcome the lack of a gold-standard is to rely on a common base of instances. To that extent, the intrinsic instance-based precision has been introduced (Thiéblin, Haemmerlé, and Trojahn, 2021). This measure consists in verifying if the sets of instances corresponding to each side of a correspondence support the same relation than those stated by the correspondence.

They also proposed to evaluate the alignment coverage relying on Competency Questions for Alignments (CQA), i.e. pairs of equivalent queries over two ontologies. The proposed measure, CQA coverage, quantifies how well a set of CQA is covered by an alignment. This measure cannot be qualified of intrinsic since it relies on CQA that are some kind of aligned knowledge. However, the authors have shown that such an approach is accurate for evaluating complex alignments.

Agreement and disagreement

Agreement and disagreement measures, introduced by (d'Aquin, 2009), allow to quantify how aligned ontologies agree or disagree. The idea is that ontologies agree when they contain compatible assertions and they disagree when they contradict each other. The approach consists in comparing the relation that exists between each pair of entities (classes, properties, instances) of a first ontology with the relation that holds in a second ontology between the images of entities obtained by a (functional) alignment.

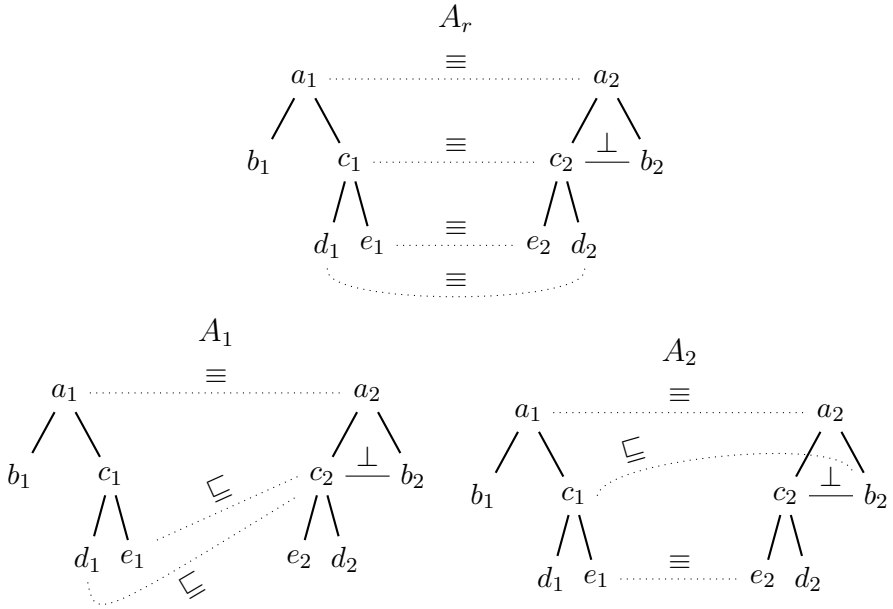


Figure 3.1: Example of a reference alignment A_r and two alignments A_1 and A_2

The degree of agreement/disagreement between two relations is given thanks to ad-hoc tables defined in (d'Aquin, 2009). The relations taken into account are subClassOf, equivalentClass, domain, range, disjointWith, type, sameAs, differentFrom, subPropertyOf, or a generic, user-defined property R, e.g. isPartOf , isAuthorOf. Even if these values follow intuitively the semantics of ontologies, they are not formally grounded on the ontology semantics.

These measures have not been designed for alignment evaluation purpose, but in an ontology retrieval perspective. However, by making explicit the role of alignments into them and generalizing this principle, this approach could be beneficial to ontology alignment evaluation. This work was the starting point from which we designed the inclusion measure presented in Section 3.3.

3.2 Extrinsic evaluation measures

We are interested in the evaluation of alignments when reference alignments are available. In this context, we have proposed new evaluation measures that take advantage of both semantic and relaxed models (David and Euzenat, 2008a). Our approach relies on a representation of alignments and relations within ontologies based on algebras of relations (Euzenat, 2008).

We start by highlighting the limits of the different models using an example and then we show how these can be fixed.

Let us take the example given in Figure 3.1 showing a reference alignment and two alignments to evaluate. The question is: which alignment among A_1 or A_2 is more correct with respect to A_r ?

Both classical and relaxed precision tell us that A_2 is better than A_1. However, if we compare the two alignments, we will see that A_1 is more correct than A_2 because all the correspondences of A_1 are correct while in A_2 the correspondence $\langle c_1, b_2, \sqsubseteq \rangle$ is

Precision	Classical	Relaxed	Semantic
A_1	1/3	1.5/3	3/3
A_2	2/3	2/3	2/3

Table 3.1: Precision values obtained by two alignments A_1 and A_2 w.r.t. A_r

Recall	Classical	Relaxed	Semantic
A_1	1/4	1.5/4	1/4
A_2	2/4	2/4	2/4

Table 3.2: Recall values obtained by two alignments A_1 and A_2 w.r.t. A_r

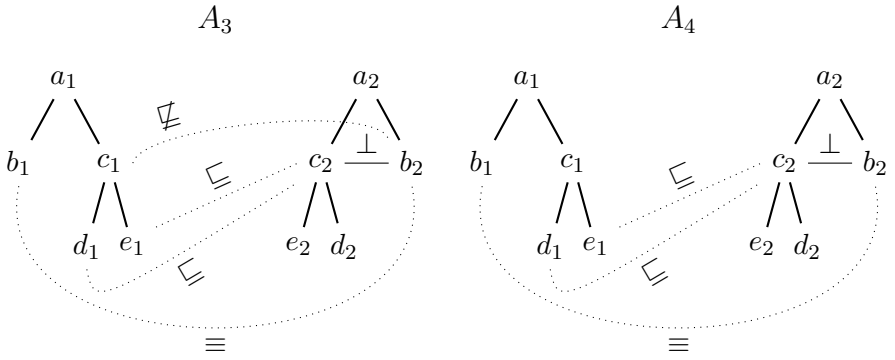


Figure 3.2: Example of two alignments A_3 and A_4 to be compared to the reference one given Figure 3.1

incorrect. Only the semantic precision is able to capture this, hence it better reflects the real correctness of alignments.

In term of recall, for the alignment A_1, neither classical nor semantic recall take into account the correspondences $\langle d_1, c_2, \sqsubseteq\rangle$ and $\langle e_1, c_2, \sqsubseteq\rangle$. However, even if these two correspondences do not belong to the reference alignment A_r, they contribute to a part of information contained in A_r and we can acknowledge that A_1 convey more information than an alignment only made of the correspondence $\langle c_1, c_2, \equiv\rangle$. Only relaxed recall is able to consider this.

Semantic measures do not necessarily evaluates in the same way semantically equivalent alignments. For instance, on Figure 3.2, A_3 and A_4 are equivalent because $\langle c_1, b_2, \not\sqsubseteq\rangle$ can be entailed thanks to the correspondence $\langle e_1, c_2, \sqsubseteq\rangle$ and the axiom $c_2 \sqcap b_2 \sqsubseteq \perp$ of the second ontology.

To overcome such problems, we have proposed to merge relaxed and semantic evaluation measures. Our approach relies on algebras of alignment relations introduced in (Euzenat, 2008) and refined in (Inants, 2016). Such an algebra allows to write any relation between concepts as a disjunction of elementary relations, for instance those of $\mathbb{A}^5$, i.e. $\Gamma = \{\sqsubset, \sqsupset, \equiv, \between, \perp\}$. Furthermore, we also take advantage of them to introduce a normal form on alignments allowing to compare them on the same syntactical basis.

Thanks to algebras of relations, proximity measures between correspondences can be defined. They are functions of the cardinality of the intersection between the rela-

O_1

	a_1	b_1	c_1	d_1	e_1
a_1	$\equiv$	$\sqsupset, \equiv$	$\sqsupset, \equiv$	$\sqsupset, \equiv$	$\sqsupset, \equiv$
b_1	$\sqsubset, \equiv$	$\equiv$			
c_1	$\sqsubset, \equiv$		$\equiv$	$\sqsupset, \equiv$	$\sqsupset, \equiv$
d_1	$\sqsubset, \equiv$		$\sqsubset, \equiv$	$\equiv$	
e_1	$\sqsubset, \equiv$		$\sqsubset, \equiv$		$\equiv$

O_2

	a_2	b_2	c_2	d_2	e_2
a_2	$\equiv$	$\sqsupset, \equiv$	$\sqsupset, \equiv$	$\sqsupset, \equiv$	$\sqsupset, \equiv$
b_2	$\sqsubset, \equiv$	$\equiv$	$\bot$	$\bot$	$\bot$
c_2	$\sqsubset, \equiv$	$\bot$	$\equiv$	$\sqsupset, \equiv$	$\sqsupset, \equiv$
d_2	$\sqsubset, \equiv$	$\bot$	$\sqsubset, \equiv$	$\equiv$	
e_2	$\sqsubset, \equiv$	$\bot$	$\sqsubset, \equiv$		$\equiv$

Figure 3.3: Closed self alignments of ontologies O_1 and O_2. An empty cell means that the relation is unknown, i.e. Γ.

tions within the correspondences. For example, let us consider the two correspondences $\langle a_1, a_2, \{\equiv\}\rangle$ and $\langle a_1, a_2, \{\sqsubset, \equiv\}\rangle$. The precision proximity is $|\{\equiv\} \cap \{\sqsubset, \equiv\}|/|\{\equiv\}| = 1$ while the recall proximity is $|\{\equiv\} \cap \{\sqsubset, \equiv\}|/|\{\sqsubset, \equiv\}| = 1/2$. This example takes the semantics of alignment into account but does not integrate those of ontologies. This can be easily fixed by using composition of relations as shown in (David and Euzenat, 2008b). The idea is to compose the relations of the first ontology with those of the alignment and finally with those of the second one. This approach is not complete but it is correct and has the advantage of offering calculable measures.

Our relaxed semantic evaluation model represents alignments as matrices where row and columns represent the named entities from the ontologies and values are sets representing the disjunction of elementary relations that hold between entities. If the relation is unknown, the disjunction is made of all elementary relations Γ.

Combined with closure, this representation acts as a normal form since the relation that holds between each pair of named entities is given. This allows to get rid of the cardinality of alignments that introduces problems when syntactically different but semantically equivalent alignments are compared.

The procedure takes as input the two self (closed) alignments of ontologies o_1 and o_2 and an alignment represented as matrix of relations. The ontology matrices are given Figure 3.3 and those of alignments are given Figure 3.4. The closure of an alignment A_x is simply obtained by doing the matrix multiplication $A_x^+ = O_1 \cdot A_x \cdot O_2$ using intersection over composition operator of algebras of relations.

Then the relaxed semantic precision and recall measures are calculated by averaging the precision, resp. the recall, proximities over all known relations, i.e. those which are different from Γ. Results are given in Table 3.3. With this relaxed semantic evaluation model, since A_1 is correct, it obtains a precision of 1. The recall is now able to take into account the two correspondences with subsumption, since without them the recall is equals to .59 only. This example also shows that the more general a correspondence is, the more influence on the precision and recall values it has. The incoherent correspondence contained in A_2 has a strong negative impact on both precision and recall. And finally, we see that both A_3 and A_4 even if they have different syntactical forms have the same precision and recall because they are semantically equivalent.

Our relaxed and semantic evaluation model solves several issues that we have illustrated previously. However, we can identify some limitations. First, it does not necessarily encode all ontology relations between entities, as for instance ternary relations such as $c_1 \sqsubseteq d_1 \sqcup e_1$. At the level of alignments, it is restricted to simple alignments with correspondences between named entities only.

If one wants to use this model in the case of complex alignment evaluation, the approach

Original alignments

A_r	a_2	b_2	c_2	d_2	e_2
a_1	$\equiv$				
b_1					
c_1			$\equiv$		
d_1				$\equiv$	
e_1					$\equiv$

A_1	a_2	b_2	c_2	d_2	e_2
a_1	$\equiv$				
b_1					
c_1					
d_1			$\sqsubset, \equiv$		
e_1			$\sqsubset, \equiv$		

A_2	a_2	b_2	c_2	d_2	e_2
a_1	$\equiv$				
b_1					
c_1		$\sqsubset, \equiv$			
d_1					
e_1					$\equiv$

A_3	a_2	b_2	c_2	d_2	e_2
a_1					
b_1		$\equiv$			
c_1		$\sqsupset, \between, \bot$			
d_1			$\sqsubset, \equiv$		
e_1			$\sqsubset, \equiv$		

A_4	a_2	b_2	c_2	d_2	e_2
a_1					
b_1		$\equiv$			
c_1					
d_1			$\sqsubset, \equiv$		
e_1			$\sqsubset, \equiv$		

Closed alignments

A_r^+	a_2	b_2	c_2	d_2	e_2
a_1	$\equiv$	$\sqsupset$	$\sqsupset, \equiv$	$\sqsupset, \equiv$	$\sqsupset, \equiv$
b_1	$\sqsubset, \equiv$				
c_1	$\sqsubset, \equiv$	$\bot$	$\equiv$	$\sqsupset, \equiv$	$\sqsupset, \equiv$
d_1	$\sqsubset, \equiv$	$\bot$	$\sqsubset, \equiv$	$\equiv$	
e_1	$\sqsubset, \equiv$	$\bot$	$\sqsubset, \equiv$		$\equiv$

A_1^+	a_2	b_2	c_2	d_2	e_2
a_1	$\equiv$	$\sqsupset$	$\sqsupset, \equiv$	$\sqsupset, \equiv$	$\sqsupset, \equiv$
b_1	$\sqsubset, \equiv$				
c_1	$\sqsubset, \equiv$	$\sqsupset, \between, \bot$	$\sqsubset, \sqsupset, \equiv, \between$		
d_1	$\sqsubset, \equiv$	$\bot$	$\sqsubset, \equiv$		
e_1	$\sqsubset, \equiv$	$\bot$	$\sqsubset, \equiv$		

A_2^+	a_2	b_2	c_2	d_2	e_2
a_1	$\equiv$	$\sqsupset$	$\sqsupset$	$\sqsupset$	$\sqsupset$
b_1	$\sqsubset, \equiv$				
c_1	$\sqsubset, \equiv$	$\varnothing$	$\varnothing$	$\bot$	$\varnothing$
d_1	$\sqsubset, \equiv$	$\sqsubset, \equiv$	$\bot$	$\bot$	$\bot$
e_1	$\sqsubset, \equiv$	$\varnothing$	$\varnothing$	$\bot$	$\varnothing$

A_3^+	a_2	b_2	c_2	d_2	e_2
a_1	$\sqsubset, \sqsupset, \equiv, \between$	$\sqsupset$	$\sqsupset, \between$	$\sqsupset, \between, \bot$	$\sqsupset, \between, \bot$
b_1	$\sqsubset, \equiv$	$\equiv$	$\bot$	$\bot$	$\bot$
c_1	$\sqsubset, \sqsupset, \equiv, \between$	$\sqsupset, \between, \bot$	$\sqsubset, \sqsupset, \equiv, \between$		
d_1	$\sqsubset, \equiv$	$\bot$	$\sqsubset, \equiv$		
e_1	$\sqsubset, \equiv$	$\bot$	$\sqsubset, \equiv$		

A_4^+	a_2	b_2	c_2	d_2	e_2
a_1	$\sqsubset, \sqsupset, \equiv, \between$	$\sqsupset$	$\sqsupset, \between$	$\sqsupset, \between, \bot$	$\sqsupset, \between, \bot$
b_1	$\sqsubset, \equiv$	$\equiv$	$\bot$	$\bot$	$\bot$
c_1	$\sqsubset, \sqsupset, \equiv, \between$	$\sqsupset, \between, \bot$	$\sqsubset, \sqsupset, \equiv, \between$		
d_1	$\sqsubset, \equiv$	$\bot$	$\sqsubset, \equiv$		
e_1	$\sqsubset, \equiv$	$\bot$	$\sqsubset, \equiv$		

Figure 3.4: Alignments and their closures. An empty cell means that the relation is unknown i.e. Γ, while $\varnothing$ represents incoherency.

	Precision	Recall
A_1	1	.78
A_2	.38	.69 (.47)
A_3	.75	.62
A_4	.75	.62

Table 3.3: Relaxed semantic precision and recall value of alignments w.r.t. A_r

consists in adding new row and/or columns in the matrices. These new rows and columns represents all class expressions that appear in complex alignments or in ontologies. In the case of evaluation campaigns, where the alignments can contain different class expressions, it would require to consider all of these expressions together, if we want to compare all alignments on the same basis. However, as previously mentioned, this requires to determine the relations holding between expressions that appear in the correspondences. This is not straightforward in the case of query languages allowing transformation functions.

All these evaluation models still rely on reference alignments. This kind of measures is useful in controlled evaluation environments such as OAEI however in real conditions, references does not exist. In the following section, we address the problem of evaluating the quality of alignments without references and we propose intrinsic evaluation measures.

3.3 Intrinsic evaluation measures

Approaches to intrinsic evaluation of alignments mostly rely on either the coherency or the conservativity principles. There are both semantic evaluation models. However, they are not very informative: there are many conservative or coherent alignments that are incorrect.

We propose to measure the quality of alignments thanks to inclusion and exclusion indexes between ontologies modulo the alignment. This is in fact a generalisation of agreement and disagreement measures proposed by (d'Aquin, 2009) that makes explicit the role of alignments in these measures and that considers the semantics of ontologies and alignments. A main difference with the aforementioned measures is that our vision of inclusion is not symmetric while the agreement is symmetric. For instance, if a first ontology asserts that B is a subclass of A and second one states that A and B are equivalent classes, according to (d'Aquin, 2009), the two ontologies partially agree. In our opinion, this notion of agreement depends on the point of view, and is directly linked to entailment: the first ontology will not fully agree with the second one while the second one will agree with first one. This justify the need of asymmetric measures of inclusion and exclusion.

The degree of inclusion of one ontology O_1 into an ontology O_2 given an alignment A is the proportion of axioms of O_1 that are entailed by the union of A and O_2. The degree of exclusion from O_1 to an ontology O_2 given an alignment A is the proportion of axioms of O_1 that are contradicted by the union of A and O_2.

At a first glance, the idea seems simple, but there are several dimensions to consider and different measures can be developed. In particular, we can distinguish the following dimensions:

- The **model** used to represent ontologies: They are usually represented as a set of OWL/DL axioms but we can also use simplifications such as the set of relations between concepts. These relations can be expressed as triples (RDF) or using some algebra of relations (Inants, 2016).

- The **basis**: The measure can be calculated on the basis of all the concepts declared in the ontologies, but it can be also restricted to only aligned ones. While the first one better reflects how an ontology is included into another one, the latter can be useful in evaluation contexts where for the same level of inclusion, a smaller alignment may be an indicator of a better precision.

- The **inclusion/exclusion checking**: the inclusion may be reduced to entailment/unsatisfiability checking, but it can also rely on predefined values such those given in (d'Aquin, 2009) or computed thanks a proximity between relations.

In the following, we investigate two families of measures. The first one models ontologies as a set of OWL axioms and check inclusion, resp. exclusion, thanks to entailment, resp. unsatisfiability. The second family represents ontologies as binary relations between concepts thanks to algebras of relations. Inclusion and exclusion checking rely on set inclusion between disjunctive relation, but can be defined thanks to proximities in the same way as for relaxed semantic measures presented in the previous Section 3.2.

Inclusion and exclusion measures based on axioms

Ontologies can be seen as a set of axioms. As shown in Figure 3.5, measures will consist in checking for each axiom of the first ontology, if it is entailed by the union of the other ontology and the alignment. For instance, the inclusion degree between o_1 and o_2 modulo A is 3/5 if the basis considers all axioms or 3/4 if it is restricted to only aligned axioms. The exclusion degree is 1/5 (or 1/4 in case of aligned axioms).

Our hypothesis, is that the measure of inclusion, when computed on the full basis, is correlated to the recall measure: the higher the recall is, the more correct correspondences are contained the alignment and then the higher is the inclusion degree. But when restricted to the aligned basis, it could be more correlated with the precision.

These measures can be good candidates for evaluating the quality of alignment if no reference is given. However they are sensitive to modifications in the alignment, for instance, if the correspondence $\langle c_1, c_2, \equiv\rangle$ is removed, then the inclusion degree drops to 1/5 because axioms containing class c_1 cannot be entailed any more. However, one can see that some part of their consequences, e.g. `SubClassOf(:e1 :a1)`, is still entailed.

In order to fix such behaviour, the closure of ontology can be considered, but it may be large, if not infinite. A compromise is to focus on the relations between every pair of concepts instead of axioms.

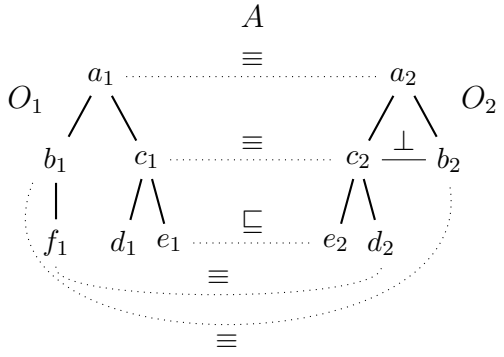
Inclusion and exclusion measures based on relations

To illustrate the indexes based on relations, we use the example of Figure 3.5. If one want to measure the degree of inclusion of O_1 into O_2 given A, she will have to compose $A.O_2.A^{-1}$. This composition gives a representation of relations between concepts of O_1 that can be deduced from the alignment A and the ontology O_2 as shown in Figure 3.6.

Since the relations are reversible, we only consider half part of matrices of Figure 3.6. The two measures are obtained by comparing the informative relations, i.e. different from Γ, from the matrix of O_1 to those of the composition. On Figure 3.6, relations in dark grey include those of the composition and are then counted as inclusions, while the one in light grey has an empty intersection with the composition and is then counted as an exclusion.

On this example, the degree of inclusion, resp. exclusion, of O_1 into O_2 modulo A is equal to 9/14, resp. 1/14, if all entities are included in the basis. If the basis is restricted to only aligned entity the denominator becomes 11.

The use of this model based on relation algebra makes the measures less sensitive to alignment than those based on asserted axioms. For instance, if $\langle c_1, c_2, \equiv\rangle$ is removed, we



O_1	$A \cup O_2$	O_2	$A \cup O_1$
`SubClassOf(:b1 :a1)`	$\models$	`SubClassOf(:b2 :a2)`	$\models$
`SubClassOf(:c1 :a1)`	$\models$	`SubClassOf(:c2 :a2)`	$\models$
`SubClassOf(:d1 :c1)`	-	`SubClassOf(:d2 :c2)`	$\not\models$
`SubClassOf(:e1 :c1)`	$\models$	`SubClassOf(:e2 :c2)`	$\not\models$
`SubClassOf(:f1 :b1)`	$\bot$	`DisjointClasses(:b2 :c2)`	$\not\models$

Figure 3.5: Axioms that are entailed by ($\models$) or incoherent with ($\bot$) the union of the alignment and the other ontology. An empty cell means that the axiom contains not aligned concept.

O_1

	a_1	b_1	c_1	d_1	e_1	f_1
a_1	$\equiv$					
b_1	$\sqsubset, \equiv$	$\equiv$				
c_1	$\sqsubset, \equiv$		$\equiv$			
d_1	$\sqsubset, \equiv$		$\sqsubset, \equiv$	$\equiv$		
e_1	$\sqsubset, \equiv$		$\sqsubset, \equiv$		$\equiv$	
f_1	$\sqsubset, \equiv$	$\sqsubset, \equiv$				$\equiv$

$A.O_2.A^{-1}$

	a_1	b_1	c_1	d_1	e_1	f_1
a_1	$\equiv$					
b_1	$\sqsubset, \equiv$	$\equiv$				
c_1	$\sqsubset, \equiv$	$\bot$	$\equiv$			
d_1						
e_1	$\sqsubset, \equiv$	$\bot$	$\sqsubset, \equiv$			
f_1	$\sqsubset, \equiv$	$\bot$	$\sqsubset, \equiv$			$\equiv$

Figure 3.6: Closed self alignments of ontology O_1 and the composition $A \cdot O_2 \cdot A^{-1}$. An empty cell means that the relation is unknown i.e. Γ

observe a decrease from 9/14 to 6/14 but in the previous model the decrease was from 3/5 to 1/5.

Interestingly, these measures can be relaxed following the approach of relaxed semantic precision and recall presented in Section 3.2. By doing so, we can generalize measures of agreement and disagreement proposed by (d'Aquin, 2009). In fact instead of relying on an ad-hoc table with the different levels of agreements and disagreement, they can be computed thanks to inclusion or exclusion degree of disjunctive relations.

As pointed out in the previous section, such a model is restricted to binary relation between concepts. For instance and contrary to the measures based on axioms, algebra of relations can not encode axioms between classes and properties such as the domain or range.

3.4 Conclusions and perspectives

This chapter focused on providing metrics for evaluating ontology alignments. In this context, we studied both extrinsic and intrinsic evaluation. While the former relies on a gold standard, the latter attempts to estimate the quality of the alignment without reference. One of its originality lays in intrinsic metrics which have rarely been studied. However, this type of evaluation is more appropriate in real-life scenarii where the ground truth is not known. Another originality is the consideration of ontology semantics and alignments in the proposed metrics.

At the extrinsic evaluation level, we merged the semantic and relaxed approaches. Using relation algebra, we have defined a kind of alignment normal form. Unlike classical metrics for precision and recall evaluation, this formalization has the advantage of giving the same score to semantically equivalent alignments, while being computable.

We also studied the intrinsic assessment of alignment. Inspired by the agreement and disagreement proposed by M. d'Aquin, we generalized them to propose inclusion and exclusion measures. The inclusion, resp. exclusion, measure quantifies the proportion of axioms from one ontology that are included in, resp. are contradicted by, the other modulo the alignment. We have also provided a model of these measures based on an algebra of relations.

Although this work still needs to be experimented, we think that it could be extended following two directions: the first one is the evaluation of complex alignments and the second one is the use of inclusion and exclusion measures for cultural knowledge evolution.

The first perspective is to explicitly extend these evaluation models to complex alignments. They have the advantage, contrary to classical precision and recall, to handle more than equivalence relation in the correspondences. However, they are still mostly restricted to correspondences matching concepts names (classes and properties). If one would evaluate more expressive alignments, i.e. complex alignments where aligned entities can be concept descriptions or SPARQL queries, then query containment as to be considered (David, Euzenat, Genevès, et al., 2018).

Cultural knowledge evolution applies concepts from biological evolution to knowledge seen as culture (Euzenat, 2017). In such a field, experiments are made with agents that use ontologies to represent their knowledge. They iteratively play an interaction game that may succeed or fail, and if the game fails then an adaptation or revision is made on their ontologies. In such a game, the agreement between agents will increase over iterations. We think that the inclusion measure, proposed Section3.3, could be used to predict the success rate or the convergence speed of experiments made in the field of cultural knowledge evolution.

The inclusion measure has been designed for alignment evaluation, but if one considers that the alignment is "the reference" one, it can be seen as hybrid ontology metric that both consider ontology content and alignment. One could expect that the more the ontologies agrees, i.e. the higher the inclusion values, the more chances the game will succeed. Conversely, the more ontologies disagree, the more chance the game will fail. Another hypothesis to test is that these measures are clues for guessing the convergence speed of success rate.

Chapter 4

Algorithms and measures for data interlinking

There are large amounts of RDF data available on the Web, in the form of knowledge graphs or as part of linked open data. Interoperability between RDF datasets largely relies on links between resources from different RDF datasets and especially links asserting the identity of resources bearing different IRIs, specified using the owl:sameAs property (Heath and Bizer, 2011). Since RDF datasets tend to be large, the automatic discovery of owl:sameAs links between RDF datasets is an important and challenging task. This task is usually referred to as data interlinking and different algorithms and tools for data interlinking have been proposed (Ferrara et al., 2011; Nentwig, Hartung, Ngonga Ngomo, et al., 2017b).

We addressed the challenge of data interlinking with a symbolic approach. Our goal was to investigate unsupervised methods for data interlinking that also minimize the required input. We started with the notion of key in RDF combined with ontology alignments and we finally designed the generalizing notion of link keys.

In databases, keys are functional dependencies that allow to identify tuples in a given relation. They are a meaningful and well-defined model for performing deduplication and identification. They have also the advantages to be easily integrated to ontologies thanks to the owl:hasKey constructor and they can be expressed as link specification used by tools such as SILK (Volz et al., 2009) or LIMES (Ngonga Ngomo and Auer, 2011).

We have designed an algorithm for extracting keys in RDF datasets. It is inspired by those used for the discovery of functional dependencies in databases (Mannila and Raiha, 1994; Huhtala et al., 1999). The main difference is that we deal with non functional properties. We have also made the distinction between different flavours of keys and introduced the notion of pseudo-keys. The quality of pseudo-key can be assessed with measures such as discriminabilty and support.

When one wants to perform data interlinking with keys, an alignment between ontologies of the datasets may be needed because datasets do not necessarily use the same vocabulary. In this case, keys have to be discovered between each dataset independently, and then only keys for which properties are aligned can be used. This may be expensive because it will require ontology matching and key extractions. To overcome such a limitation, we have introduced the notion of link key that can been seen as keys defined on the intersection of datasets.

4.1 Related work

Data interlinking is reminiscent of the task of record linkage in databases (Christen, 2012), but it is applied to RDF data eventually described with RDFS/OWL ontologies.

The first data interlinking frameworks to be proposed are RDF-AI (Scharffe, Liu, et al., 2009), KnoFuss (Nikolov, Uren, et al., 2008), SILK (Volz et al., 2009) and LIMES (Ngonga Ngomo and Auer, 2011). They allow someone to define link specifications and process them to generate owl:sameAs links between two RDF datasets. Link specifications are similarity rules specifying the conditions that instances must fulfil in order to be equal. They express the properties to compare, the transformations applied to them, the similarity measures to use for comparing pairs of property values, the aggregation functions for combining several similarity values, and the thresholds beyond which two values are considered equal.

Contributions around these frameworks include languages to express link specifications and strategies for optimizing the generation of links both in terms of time efficiency and scaling. SILK makes use of blocking strategies that has been improved in (Isele, Jentzsch, et al., 2011), while LIMES takes advantage of triangular inequality in the Euclidean space to reduce the number of comparisons, and optimize execution by rewriting link specifications (Ngonga Ngomo, 2014).

Beside processing link specifications, work has been carried out on the semi-automatically discovery of link specifications from data. This has been addressed mainly by using machine learning techniques such as been done in EAGLE (Ngonga Ngomo and Lyko, 2012), ActiveGenLink (Isele and Bizer, 2013), COALA (Ngonga Ngomo, Lyko, and Christen, 2013). These techniques are mainly supervised, so they require a set of reference links or user feedback. They are not able to automatically align properties and to select the classes to compare, i.e. part of the specifications. An exception is the KnoFuss system which is able to learn similarity rules in an unsupervised way using a genetic algorithm (Nikolov, d'Aquin, et al., 2012). These propositions mostly focus on discovering the best similarity and threshold to find links. Furthermore, none of them consider the semantics of ontologies describing instances.

A different approach to data interlinking take advantage of idea of functional dependencies transposed to RDF. In OWL, inverse-functional properties or more generally the owl:hasKey axiom are particular kinds of functional dependencies. Together with an alignment between ontologies, they allow to interlink data (Atencia, David, and Euzenat, 2021). L2R (Saïs, Pernelle, et al., 2007) has been the first logical method for interlinking data that takes advantage of functional properties, inverse-functional properties and disjointness axioms. This method has been complemented by the N2R numeric approach to deal with syntactic variations in values (Saïs, Pernelle, et al., 2009). In (Hogan et al., 2012), a scalable method to data interlinking has been proposed. It uses a subset of OWL 2 RL rules related to the semantics of owl:sameAs, functional properties, inverse functional properties and max-cardinality restrictions with value one. These methods are automatic and unsupervised because they neither need linkage rules such as LIMES or SILK nor training set of links. However, they rely on declarative axioms that are not necessarily available in ontologies. Hence methods for discovering them from data are needed.

A first method for discovering inverse-functional properties in RDF has been proposed by (Song and Heflin, 2011). They relax the notion of inverse-functional properties using the notions of coverage and discriminability of a property. The coverage of a property is defined as the ratio of the number of instances of a class having that property to the total number of instances of that class. The discriminability of a property is the ratio of

the number of distinct values for the property to the total number of instances having that property. Syntactic variations in values are handled with string matching techniques. This interlinking method still need an alignment between properties (and classes) as input. In (Hogan et al., 2012), thanks to an inverse cardinality index, quasi inverse-functional properties can also be discovered.

Inverse-functional properties are superseded by keys which are a more general way to perform logical data interlinking. As for inverse-functional properties, keys are not necessarily provided with ontologies and need therefore be discovered from data. The problem of discovering keys has been heavily studied in the database field (Huhtala et al., 1999; Sismanis et al., 2006). But in the semantic web, the non functionality of properties makes the problem different. A first approach to discover keys in RDF, named KD2R, has been proposed in (Symeonidou, Pernelle, et al., 2011). KD2R is based on the Gordian algorithm (Sismanis et al., 2006) that uses a depth-first search strategy. This approach does not allow exceptions. We have then proposed an alternative method that allows keys with exceptions, called pseudo-keys (Atencia, David, and Scharffe, 2012). We have shown that the semantics of keys discovered by KD2R (and those of OWL2) differs from ours (Atencia, Chein, et al., 2014). Several other key discovery methods have been published (Symeonidou, Armant, et al., 2014; Soru et al., 2015).

Having keys between two datasets is not sufficient for interlinking data. They have to be compatible, in the sense that their properties have to be aligned. Key-based approaches typically extract keys from RDF data sets, select, and combine them with ontology alignments for interlinking (Symeonidou, Armant, et al., 2014; Achichi, Cheatham, et al., 2016; Farah et al., 2017; Atencia, David, and Scharffe, 2012). Unlike (Symeonidou, Armant, et al., 2014), the key-based approaches to data interlinking proposed in (Achichi, Ellefi, et al., 2016; Farah et al., 2017) aim to discover S-keys that hold not only in the source dataset, but in both source and target datasets. However, it is assumed that the datasets are described using the same vocabulary, possibly resulting from merging different ontologies with an alignment, again composed of equivalence correspondences only.

Most of the approaches to data interlinking need input such as alignments, ontology axioms (functional properties, inverse-functional properties, keys), link specifications and/or training set of links. My work focused on in designing a logical interlinking approach minimizing the required input. To that extent, we have developed the notion of link keys. The discovery of link keys requires only two RDF datasets as input without alignments, ontologies, or training set of links.

4.2 Key and pseudo-key discovery

Our work on data interlinking has led us to study key extraction from RDF graphs. Indeed, even the OWL language allows declaring key axioms, it is rare to find such predicates in ontologies. Furthermore, the variable quality of web data makes it necessary to tolerate some exceptions, which led us to introduce the notion of *pseudo-key*.

In (Atencia, David, and Scharffe, 2012), we proposed an algorithm to extract keys and pseudo-keys from RDF data. We showed experimentally that the algorithm scales to large datasets such as DPBEDIA (13.8 M triples) where more than 9000 keys (and pseudo-keys) were extracted in less than 3 hours. We also demonstrated on this example that pseudo-keys can be used for error detection.

In (Atencia, Chein, et al., 2014), we have with other colleagues investigated two kinds of keys since RDF properties are multivalued, contrary to relational attributes, which are

mono valued. They mainly differ on whether the Open World Assumption or the Closed World Assumption when considering the properties within the key. If a set of properties form an S-key for a class, it is enough that two instances of the class share *one* value for each of the properties of the key to infer that they are the same, e.g. email property for the AssistantProfessor class. But if the properties form an F-key then the instances must share *all* values, e.g. hasPoem property for the PoemAnthology class because two different poem anthologies may have a poem in common but will unlikely contain exactly the same poems. The keys considered in (Atencia, David, and Scharffe, 2012) are F-keys contrary to approaches such as (Symeonidou, Armant, et al., 2014; Achichi, Ellefi, et al., 2016; Farah et al., 2017) that extract S-keys. The extension to S-keys of the pseudo-key algorithm has also been implemented.

Principle of the pseudo-key extraction algorithm

The complexity of finding keys in an RDF graph is polynomial in the number of subjects, but exponential in the number of predicates. Our approach to extract pseudo-keys adopt the same breadth first search strategy as the functional dependencies discovery algorithm TANE (Huhtala et al., 1999). It explores the lattice of the power set of properties starting from singleton property sets. For each set of properties, it builds the partition of instances where each equivalence class contains the instances having the same set of values for the properties. If the ratio of singleton within the partition, named discriminability, is greater than a fixed threshold, it is then a pseudo-key. There are two advantages in this approach: it facilitates the pruning of the search space and reduces the cost for computing the partitions of instances. The pruning rule is the following: If a given set of property is a pseudo-key or its support, i.e. the ratio of instances covered by the partition, is too low or it contains functional dependency, then all its super sets of properties are discarded. This pruning strategy also ensures that only minimal keys will be generated.

This notion of pseudo-key relies on the measures of support and discriminability. They allows to consider the specificities of knowledge graphs such as the open world assumption and the use of support threshold can be useful because properties are not necessary instantiated for each individual.

Contrary to the TANE algorithm, the use of support threshold can be useful because properties are not necessary instanciated for each individual. But, in counterpart, the optimization consisting in stripping partition (Huhtala et al., 1999) (removing singleton sets) can not be used. Finally, since the goal of the algorithm is to find keys only, there is no need to test exhaustively all the functional dependencies.

Interlinking with keys

As mentioned previously, keys alone are not sufficient for interlinking datasets expressed with different ontologies. Our key based interlinking is an extension of the process proposed by (Scharffe and Euzenat, 2011). As shown on Figure 4.1, it takes as input the two datasets, their respective ontologies and an alignment between them. In a first step, pseudo-key discovery is performed on both dataset separately. Then from the alignment and the two sets of extracted keys, aligned keys for the two dataset are selected. Finally, these aligned keys can be transformed to SPARQL queries for generating links. They can also be a good basis for generating link specifications and used with tools like SILK (Bizer et al., 2009), LIMES (Ngonga Ngomo and Auer, 2011), KnoFuss (Nikolov, Uren, et al., 2008), or RDF-AI (Scharffe, Liu, et al., 2009).

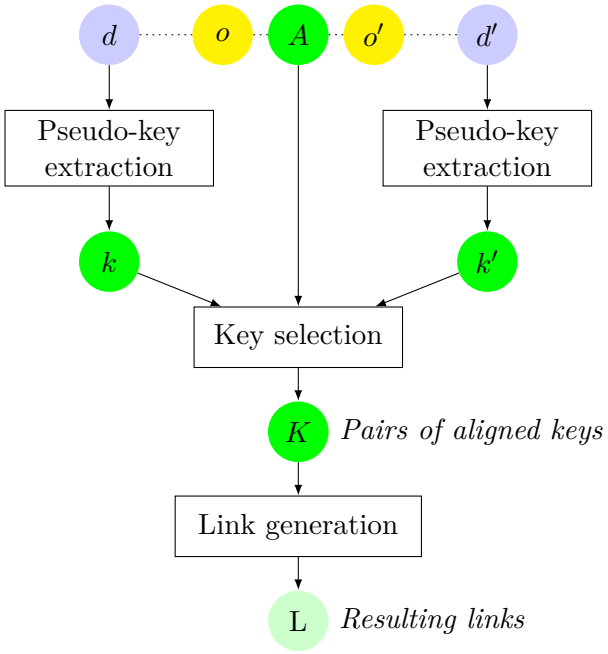


Figure 4.1: Key-based data interlinking workflow

This is a generic workflow that can be adapted and optimized for particular interlinking tasks. For instance, one could take advantage of alignments to optimize the pseudo-key extraction process: in fact this can be performed only on properties that are aligned. The generation of link specifications could also take advantage of the discriminability values of pseudo-keys: we could imagine for instance an iterative process that starts by applying the best keys and then iteratively apply the other on the remaining instances.

Application to error detection

Experimenting the pseudo-key extraction algorithm led us to consider another application: the detection of errors in a dataset. When slightly relaxing the notion of keys by decreasing the discriminability threshold in order to detect pseudo-keys, we see appearing keys that are valid for most instances but are not keys for a small number of instances. Observation of these instances reveals the presence of duplicates or errors in the dataset. In order to find errors, we transform pseudo-keys found in that way into SPARQL queries to retrieve only instances having the same values for the properties of the key. We then use the query results as a basis for error correction. This workflow is illustrated Figure 4.2.

We have applied this method on the 244 classes of the DBPedia dataset. We give below an example of pseudo-keys obtained for the class dbo:Person computed with a minimal support $\lambda_s = 0.2$ and a discriminability threshold $\lambda_d = 0.999$. Table 4.1 shows computed keys and their support.

The first row of this table indicates that there exist persons born on the same day who also died on the same day, which is not impossible but statistically rare. A verification can be performed by transforming pseudo-keys into SPARQL queries and executing them on the dataset.

The query must check what the resources have the same values for properties in the key for the given class.

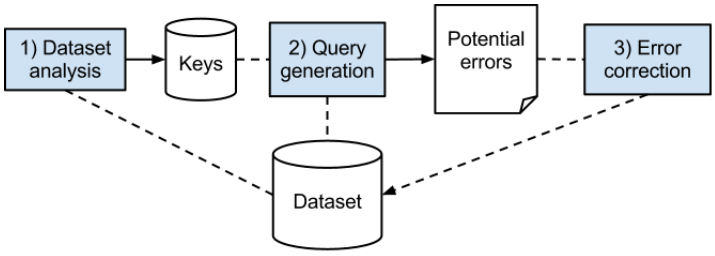


Figure 4.2: Error detection using keys: workflow

Properties of the key	Support
http://dbpedia.org/ontology/deathDate http://dbpedia.org/ontology/birthDate	0.203
http://dbpedia.org/ontology/deathDate http://dbpedia.org/ontology/deathPlace	0.216
http://xmlns.com/foaf/0.1/name http://dbpedia.org/ontology/birthPlace	0.442
http://xmlns.com/foaf/0.1/surname http://purl.org/dc/elements/1.1/description	0.459
http://dbpedia.org/ontology/deathPlace http://dbpedia.org/ontology/birthDate	0.480

Table 4.1: Key detection for the class DBPedia:Person.

We obtain the following query for the key (dbo:birthPlace, dbo:deathPlace):

```
SELECT DISTINCT ?x ?y
WHERE {
   ?x dbo:deathDate ?dp1;
   dbo:birthDate ?dp2;
   rdf:type dbo:Person.
   ?y dbo:deathDate ?dp1;
   dbo:birthDate ?dp2;
   rdf:type dbo:Person.
MINUS {
   ?x dbo:deathDate ?dpx1;
   dbo:birthPlace ?dpy1 .
   ?x dbo:deathDate ?dpx2 ;
   dbo:birthPlace ?dpy2 .
   FILTER (?dpx1=?dpy1)
   FILTER (?dpx2=?dpy1)
}
FILTER (?x!=?y) }
```

In this example, the MINUS query pattern is not required because dbo:birthDate and dbo:deathDate are single-valued properties. But in case of multivalued properties this operator is needed.

Manual analysis of the query results[1] shows the 124 instances pairs returned by the query in fact correspond to diverse types of errors in the dataset. The first kind of errors arises when two resources exist for describing a same object, for example dbpedia:Louis_IX_of_France__Saint_Louis___1 and dbpedia:Louis_IX_of_France

[1]Query executed on the DBPedia SPARQL endpoint http://dbpedia.org/sparql

Class	duplicate	misclassification	other
dbpedia:Person	31	75	16

Table 4.2: Repartition of errors in the DBPedia class dbo:Person

A second kind of errors seems to be caused by the infobox extraction process when generating DBPedia. These errors most of the time lead to resource misclassification problems. For example: dbpedia:Timeline_of_the_presidency_of_John_F._Kennedy is classified as a person although it is in fact a timeline.

Finally, a third kind of errors come from Wikipedia inconsistencies between the infobox and the article[2] or from documents from which these articles were informed.[3]
Table 4.2 below shows error distribution for the class dbo:Person.

This method can be reproduced on any dataset without any prior knowledge of the data.

Conclusion

One solution to data interlinking relies on keys. However such keys are usually not provided and thus they need to be discovered from data. We have proposed an algorithm for computing keys and pseudo-keys in RDF graphs. The algorithm is efficient, even on large datasets thanks to pruning techniques based on measures of support and discriminability.

We have demonstrated the benefits for such an algorithm in two applications: datasets interlinking and duplicates and error detection in data. Error detection allows to efficiently detect duplicates or correct errors on DBPedia persons.

4.3 Link key discovery

In the previous section, we showed that data interlinking can be done using key pairs linked by alignments. However, it may happen that no alignment is available, or there is no common key between datasets. Following our key-based approach to data interlinking, we sought a way to minimize the required input as much as possible. This gave rise to the notion of link key that generalise the combination of keys and ontology alignments for data interlinking (Euzenat and Shvaiko, 2013; Atencia, David, and Euzenat, 2014a).

A link key is composed of two sets of pairs of properties and a pair of classes. The two sets allow for distinguishing between the intersection (shortened in-) and equality (shortened eq-) parts of a link key, as it is as been done for keys (Atencia, Chein, et al., 2014). An example of a link key is:

$$\{\langle \mathsf{auteur}, \mathsf{creator}\rangle\}\{\langle \mathsf{titre}, \mathsf{title}\rangle\} \text{ linkkey } \langle \mathsf{Livre}, \mathsf{Book}\rangle$$

This link key states that whenever an instance of the class Livre has the same values for the property auteur as an instance of the class Book has for the property creator and they share at least one value for their properties titre and title, then they denote the same entity.

[2]See for example http://dbpedia.org/resource/Phromyothi_Mangkorn and http://dbpedia.org/resource/Kraichingrith_Phudvinichaikul

[3]See for example http://dbpedia.org/resource/Merton_B._Myers and http://dbpedia.org/resource/William_J._Pattison and the footnote at the end of these articles.

D		D'	
$\langle a_1, p_1, v_1\rangle$	$\langle a_2, p_2, v_4\rangle$	$\langle b_1, q_1, v_1\rangle$	$\langle b_2, q_2, v_2\rangle$
$\langle a_1, p_2, v_2\rangle$	$\langle a_2, p_3, v_5\rangle$	$\langle b_1, q_2, v_2\rangle$	$\langle b_2, q_2, v_4\rangle$
	$\langle a_2, p_1, v_3\rangle$	$\langle b_2, q_1, v_1\rangle$	$\langle b_2, q_3, v_5\rangle$

Table 4.3: Two sets of triples representing datasets D and D'

	$\langle p_1, q_1\rangle$	$\langle p_2, q_2\rangle$	$\langle p_3, q_3\rangle$
$\langle a_1, b_1\rangle$	$\times$	$\times$	
$\langle a_1, b_2\rangle$	$\times$	$\times$	
$\langle a_2, b_2\rangle$		$\times$	$\times$

Table 4.4: Context build from data given on the example of Table 4.3.

Unlike keys, this link key could be directly used to interlink the books of English and French libraries without the need of any ontology alignment.

From their introduction in (Atencia, David, and Euzenat, 2014a), we have studied different facets of link keys: their extraction, selection, combination and semantics.

(Atencia, David, and Euzenat, 2021) studied the semantics of link keys and their relations with keys. It especially shows that link keys cannot be reduced to pairs of keys related by alignment. In this work, we define three different types of link keys: weak, plain and strong link keys. They all allow to find links between two datasets, but they differ on whether they allow the existence of different resources (duplicates) satisfying the key conditions within each of the datasets: weak link keys allow them; plain link keys allow them only among the non-linked resources; strong link keys disallow them all.

Extraction algorithms

All algorithms for link key extraction focus on discovering weak link keys. The first algorithm (Atencia, David, and Euzenat, 2014a) only focused on weak in-link keys but as been then extended for discovering also weak eq-link keys.

The original algorithm consisted in extracting all the sets of pairs of properties for which instances share at least one value and that are maximal for at least one instance (Atencia, David, and Euzenat, 2014a). These set of pairs of properties has been named link key candidates and are evaluated thanks to measures discussed in Section 4.3. For instance, from the datasets given Table 4.3, the algorithm builds a context given Table 4.4 from which link key candidates $\{\langle p_1, q_1\rangle, \langle p_2, q_2\rangle\}$ and $\{\langle p_1, q_1\rangle, \langle p_2, q_2\rangle\}$ can be directly derived.

Thanks to commonalities with functional dependencies extraction in formal concept analysis (FCA), we formalized the extraction of link keys with this framework (Atencia, David, and Euzenat, 2014b), and then extended it to deal with non functional properties (Atencia, David, Euzenat, et al., 2020). Link key discovery based on FCA yields to a lattice in which the intent of concepts are the link key candidates and the extents are the set of links generated by the link key candidate. For instance, from the context given Table 4.4, FCA builds the lattice given Figure 4.3. Apart the top and bottom concepts, we can see that the link key candidate represented by the concept with intent $\langle p_2, q_2\rangle$ was not present with the original algorithm. The set of link key candidates extracted by FCA

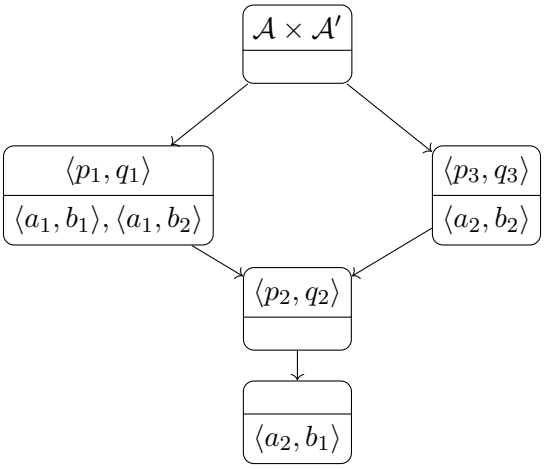


Figure 4.3: Lattice obtained from the formal context given Table 4.4.

is a super set of those obtained with the original algorithm since the resulting lattice is closed by intersection.

Until now, we have studied the extraction of independent link keys. However, a link key may be composed of some object properties that also refer to instances that have to be linked. For instance, on the example of Figure 4.4, we can see that a link key candidate composed of the pair $\langle$o1:owner, o2:ownedBy$\rangle$ requires the instances of both class o1:Person and o2:Inhabitant to be identified and hence depends on some other link key.

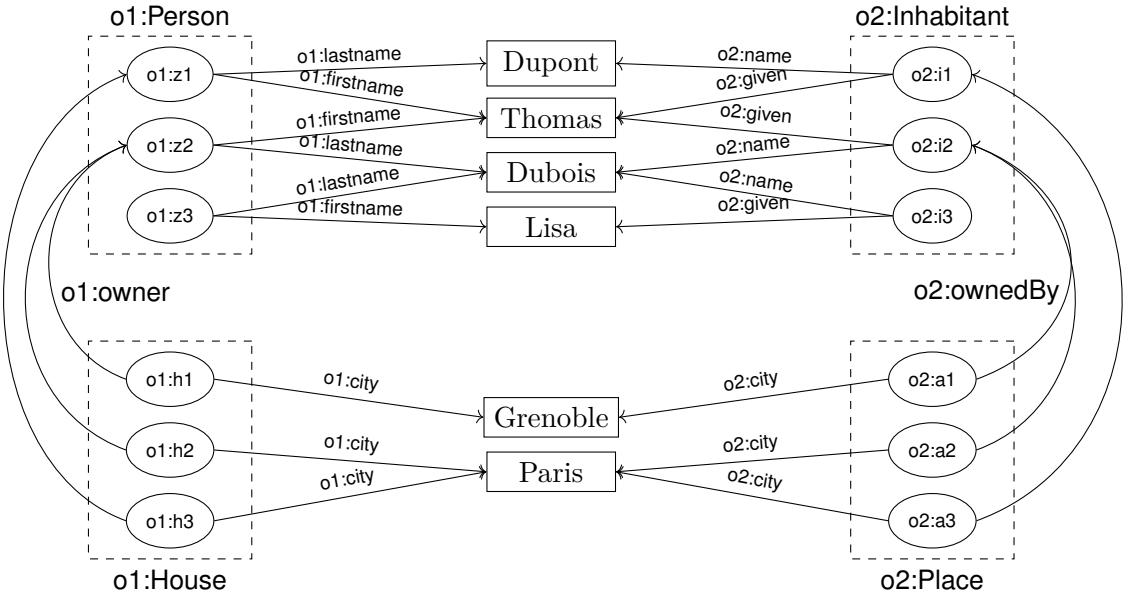


Figure 4.4: Two datasets representing instances of class House, resp. Place, that are in relation through the owner property, resp. ownedBy, with instances of class Person, resp. Inhabitants.

For this reason, we have also formalized, in (Atencia, David, Euzenat, et al., 2020), the extraction of dependant link keys within the framework of Relational Concept Analysis (RCA (Rouane-Hacene et al., 2013)) where relations between objects can be explicitly

handled.

All the algorithms presented so far discover link key for a given pair of classes. If no classes are given, then the discovery is performed on the whole graphs without considering particular classes. In (Abbas et al., 2020), we have proposed a first formalization of link key discovery thanks to Pattern Structures, an extension of FCA dealing with complex data such as numbers, trees, and graphs (Ganter and Kuznetsov, 2001). In this work, all components of link keys are represented: not only the pairs of properties, but also the class expression covering all the instances of the descriptions. These class expressions are disjunctions of conjunctions of named classes. This allows to build a lattice where intents represent fully a link key candidate, i.e. the sets of pairs of properties and also the pair of class expressions associated to the link key.

Evaluation measures

Link key extraction algorithms are exhaustive and discover all link key candidates that hold on a given pair of datasets. Among all these candidates not all of them may be valid link keys because they can either be too general/specific or be due to chance. Thus criteria for selecting only the best candidate are needed.

We follow the idea that the quality of link key candidates only depends on the set of links that it generates. When reference links are available, the quality of a link key can be naturally evaluated thanks to precision and recall. However, this is only possible in controlled evaluation scenarii, and in general the set of reference links does not exist. Hence we have made, again, the distinction between two cases: supervised when some sample of links are available, and unsupervised when there is no link.

In the supervised scenario, we assume the existence of a partial reference sets of owl:sameAs and owl:differentFrom links. Then, the estimators of precision and recall can be computed by restricting the computation to links that are common to the link key candidate and the partial reference. When only owl:sameAs links are given, a reasonable solution is to generate owl:differentFrom links by making the Unique Name Assumption on the instances linked by owl:sameAs.

In the unsupervised case, when no reference links are given, we have defined the measures of discriminability and coverage (Atencia, David, and Euzenat, 2014a). These measures make the Unique Name Assumption on each of the datasets. Then the discriminability measures how close is the set of links from a one-to-one mapping and the coverage measures the ratio of instances that are covered by the link key candidate. On the lattice of link key candidates, these measures are monotonic/antimonotonic: the more the link key candidate is general, the higher will be the coverage, and the more specific the link key is the higher will be the discriminability. However, in the case of a lattice of link key candidates with their associated pairs of classes, the coverage normalised by the size of the class is not monotonic any more (Abbas et al., 2020).

In (Atencia, David, and Euzenat, 2014a), we experimentally observed that these measures select the best candidates in both the supervised and non supervised case. They are also robust to mistakes in the data sets and sample links. This setting is well suited for finding one-to-one sets of links.

Combining link keys

One single link key, even the best one, may not be enough to discover all links in certain data sets. This is simply the case of data sources covering different concepts. This may

also be useful if the data related to a particular class is the result of aggregating different sources that use different properties: there may be different ways to generate links. Thus, instead of selecting one single best link key candidate, it could be worth selecting the best combination of link key candidates, as it is already done for other link specifications (Sherif et al., 2017).

We addressed the specific problem of extracting boolean combinations of link key candidates from two RDF data sets (Atencia, David, and Euzenat, 2019). We defined conjunction and disjunction of link keys in terms of their generated links and we shown that conjunction does not generate any link that single link key candidates do not generate already. So we focus on extraction of disjunctions of link key candidates. This is challenging because of the large number of non redundant disjunctions of link key candidates (potentially 2^n where n is the number of link key candidates).

Since an exhaustive enumeration of all non redundant disjunctions of link key is prohibitive, we proposed two strategies for searching them. Both strategies represent the search space as a lattice of antichains of link key candidates.

The *top-k strategy* selects the top-k candidates according to some evaluation measure and then performs an exhaustive enumeration of disjunction on this selection. This assumes that the best disjunctions are those which only contain the best link key candidates.

The *expand-best strategy* performs a best-first search. It explores the disjunctions from the best individual link key candidates and by iteratively replacing the best disjunction by its expansion, i.e. the set of disjunction obtained by adding another individual link key candidate. At each step, a disjunction is selected only if it is better than those explored thus far. The process stops after x iterations without any improvement. It assumes that the better a disjunction is, the more chances that it can produce better disjunctions.

We evaluated these two strategies experimentally on eight different tests sets. Overall, our hypothesis was confirmed: disjunctions of link keys bring an improvement to data interlinking with respect to single link keys. The experimental results show that the top-10 strategy always allows to find a disjunction better than the best single link key candidate. In addition, the expand-best strategy always generates longer disjunctions than the top-10 strategy. Consequently, the expand-best strategy favours recall over precision. Furthermore, top-10 scales better than expand-best.

We also observed that the harmonic mean of discriminability and coverage, that we use for selecting disjunction, was not optimal in the sense that some generated link key candidates are better in terms of F-measure than those selected by our measure. This is especially true if the data sets are very different in size (number of instances) and when the target link set is far from a one-to-one mapping.

4.4 Conclusions and perspectives

This chapter summarizes the main contributions that we made on data interlinking. We have started with the extraction of pseudo-keys from RDF data and we have generalized them by proposing the notion of a link key. We have studied link keys thoroughly: their discovery from data, their evaluation and their semantics.

We have defined three semantics of link keys: full, plain and weak link keys. We have compared them with the corresponding key notions and show that they generalize pair of keys related by alignment.

On the discovery side, we have developed a first algorithm that extracts link key candidates from data. We have then proposed an FCA model for their extraction, and

have extended it, with RCA, to the discovery of interdependent link keys. Since these algorithms do not identify the pair of classes associated to a link key, we have extended the formalization with pattern structure in order to also extract their associated class expressions.

The quality and validity of extracted link key candidates has to assessed. We considered both supervised and unsupervised evaluation. In particular, we have proposed metrics of link key coverage and discriminability for unsupervised evaluation. Experimental results have shown that coverage and discriminability are good and robust quality indicators.

When data are noisy, extraction produces many similar variations of link key candidates. If one wants to analyse manually the set of candidates, using only discriminability and coverage to rank them is not satisfactory because they could be a lot of similar candidates. We have thus studied how to reduce the set of candidates. A first proposition has been to detect those that have the same closure of links w.r.t. owl:sameAs semantics, thanks to partition pattern structures. Experimental results show that this does not reduce sufficiently the number of candidates. Recently, we have relaxed the link set equality in order to select representative candidates. Our method consists in selecting medoids within clusters obtained by hierarchical agglomerative clustering.

We are convinced that link keys present significant advantages for data interlinking. First, they can be extracted using unsupervised algorithms that minimize the input: only a pair of RDF graphs is required, there is no need to provide correspondences between classes or properties, there is no need to choose a particular distance to compare values. Link keys are a symbolic model that can be interpreted at the difference of embeddings or neural network models. We have defined a semantics allowing to reason with it: links are consequence of them, their satisfiability and consistency can be checked (Atencia, David, and Euzenat, 2021).

This work can be pursued and extended in several directions. First, we plan to generalize extraction algorithms to the different semantics of link keys. So far, we focus on extracting weak link key candidates. But one could generalize our extraction algorithm to the other semantics of link key thanks to partition pattern structures and their extension to tolerance relation (Baixeries et al., 2018). In fact, links generated by a link key candidate form a tolerance relation over the instances (of both datasets). It can be shown that the tolerance relation of weak link key includes those of the corresponding plain link key which itself include that of full link key. We can then define a sup-preserving map on the link key candidate lattices that allows to identify the different kinds of link keys. Furthermore, modeling link key extraction with tolerance relation and pattern structure would allow us to integrate similarities between properties values.

Another application of our results is the cross dataset link prediction (or knowledge base completion). Knowledge base completion automatically predicts missing facts by exploiting information already present in the knowledge base. In the recent years, many contributions are dedicated to that task. Most of them are based on knowledge graph embeddings (Wang et al., 2017). Some more recent works address this challenge using symbolic approaches based on rules (Meilicke, Chekol, et al., 2019) or (Ferré, 2021). We think that FCA-based method to link key candidate discovery could complement rules based approaches to knowledge completion by exploiting cross-dataset knowledge. Since link key candidates are FCA concepts, minimal generators can be enumerated and implications or association rules can be computed from them (Lakhal and Stumme, 2005). Such a rule could have the form:

$$\{\langle \mathsf{auteur}, \mathsf{creator}\rangle\}\{\langle \mathsf{titre}, \mathsf{title}\rangle\} \text{ linkdep } \langle \mathsf{Livre}, \mathsf{Book}\rangle \rightarrow \langle \mathsf{annee}, \mathsf{year}\rangle$$

stating that whenever an instance of the class Livre has the same values for the property auteur as an instance of the class Book has for the property creator and they share at least one value for their properties titre and title, then they share a value for their properties annee and year. This principle, that shares similarities with matching dependencies in databases (Song and Chen, 2009), could be used to complete a graph with the data of another one.

Chapter 5

Conclusions and perspectives

5.1 Summary of contributions

My work mainly addressed the challenge of dealing with knowledge heterogeneity in the semantic web. My goal has not been to reduce heterogeneity but to facilitate knowledge exchange in such an environment. Indeed, I believe that it is important that knowledge be diverse and varied in order to better adapt to different problems. To fill the gap between knowledge structures, we rely on "glue knowledge": alignment between ontologies, and links and link keys between instances.

My contributions mainly focus on measures for the comparison of knowledge structures, but I have also developed techniques that extract new knowledge by comparing structures. In particular, I am interested in ontologies, alignments and instances.

Ontologies. Different ontologies can share similar concepts. We have addressed the comparison of ontologies by proposing and evaluating distance measures between ontologies. We considered two spaces: the ontology space and the alignment space. We proposed original measures grounded on various bases such as relational structure, ontology semantics, or natural language. From such measures, it is possible to be aware of the redundancy or gaps lying behind heterogeneity. We have shown through experiments that the measures are relevant and correlated even if they do not rely on the same bases.

Alignments. The comparison of alignments is mainly used for the evaluation of ontology matchers. We have studied both extrinsic evaluation, which uses a gold standard, and intrinsic evaluation, which is based only on the ontologies and the alignment to be evaluated. In the framework of extrinsic evaluation, our approach combines both semantic and relaxed evaluation models. These measures take into account the specificity of the alignments and make it possible to overcome the limits of the classically used evaluation model. Very few work have addressed the difficult case where no gold standard exists. In this intrinsic evaluation context, we have introduced inclusion and exclusion measures allowing to guess the quality of alignments. In both cases, we took advantage of the semantics of aligned ontologies and proposed original measures grounded on algebra of relations.

Instances. The same entity can be represented in several graphs using different ontologies. The task of identifying instances across ontologies is called data-interlinking. We developed symbolic approaches to data interlinking that take the semantics into account. We have introduced the notion of pseudo-key in RDF graphs and we have then generalized it to link keys. While pseudo keys are defined on one data set, link keys are cross data sets axioms. In both cases we have developed algorithms for their discovery from data

and design quality measures ordering them according to their accuracy. We have shown experimentally that our discovery algorithms can scale to large datasets and give accurate results.

5.2 Perspectives

My research has targeted so far the comparison of static knowledge structures. However, knowledge needs to evolve in response to changes that happen in the modelled domain or for correcting errors. Evolution of knowledge encompasses many different challenges such as ontology evolution and versioning (Flouris et al., 2008; Zablith et al., 2016), ontology learning (Buitelaar et al., 2005; Lehmann and Hitzler, 2010), or knowledge graph refinement (Paulheim, 2017).

Most of the methods for knowledge evolution focus on a particular structure: ontology, instances, alignment, etc. Based on our expertise in ontology alignments and data linking, we want to study the co-evolution of knowledge. We think that the dynamics and evolution of knowledge relies on the relations between different representations. Moreover evolving one side without caring of the others is prone to bring incoherence.

In particular, we plan to address the following research questions:

1. How can ontologies enrich mutually while maintaining their diversity? In particular, it would be a question of studying how some axioms of an ontology can be integrated in another one without merging them. Indeed, two ontologies on the same domain can have divergences because they deal with different applications.

2. How can ontologies evolve using incremental learning strategies? The idea is to see how changes on instances can influence the axioms of an ontology. Extending our work on key extraction, we aim at discovering other types of axiom patterns and inducing changes by analysing how the quality indices of extracted axioms evolve over time.

3. How to measure knowledge evolution globally? The design and evolution of ontologies like those of DBpedia, Yago, Schema.org have influenced each other. Following the approach of evolution of species, we aims at providing metrics that help to understand this co-evolution and allow to build phylogenetic trees of ontologies.

All these questions are related to the research project of the mOeX team that ambitions to understand and develop general mechanisms by which a society of agents evolves its knowledge. Thanks to the combination of cultural evolution methods, multi-agent simulation and knowledge representation, the team studies how agents can adapt and evolve their knowledge, and if evolution can preserve knowledge diversity.

My approach to address the aforementioned questions will not necessarily be based on multi-agent simulation of games, but we are interested by the principle of continuous and smooth evolution developed by cultural knowledge evolution. This is an opportunity to study if observations made from simulation are supported by existing knowledge and how adaptation mechanisms studied in this field can be applied to real knowledge graphs. More generally, we are interested in symbolic knowledge learning using few examples but based on existing knowledge to validate or even try to perform some analogical reasoning.

Mutual enrichment of ontologies

Two ontologies of the same domain can express different knowledge and this diversity allow them to be mutually enriched. Indeed, there are several reasons to preserve diversity between knowledge representations and not to merge them. Diversity allows a better resilience to changing needs. Moreover, software based on ontologies of the same domain can have different perspectives and thus different conceptualizations of knowledge.

When enriching ontologies, it is difficult to find a trade-off between introducing new knowledge without inconsistencies and resolving inconsistencies without removing the added knowledge. In order to address this problem, we need to be able to identify precisely differences between ontologies and integrate those that are compatible with their respective existing knowledge. Simple alignments, as considered by the vast majority of works, are limited to the identification of common concepts and lack expressiveness to accurately identify differences between ontologies. We therefore need complex alignments that can express the description of a concept of one ontology from the vocabulary of the other.

There exists relatively few methods for discovering complex alignments between ontologies (Thiéblin, Haemmerlé, Hernandez, et al., 2020). We propose to exploit the duality between concept description and their extension because they offer more guarantee regarding consistency. To that extent association rules based methods (David, Guillet, et al., 2007; Zhou et al., 2019) extracting alignments pattern in the form $\forall x, A(x) \sqsubseteq B(x) \wedge C(x) \wedge ...$ are suitable.

The approach will consist in extracting axioms from ontologies and complex alignments. Some of them can be directly derived from complex correspondences but one can also integrate aligned axioms by following the inclusion measure approach presented in Section 3.3. The choice of axioms should preserve both consistency and diversity. If either revision should be applied because of inconsistency, we should be sure that diversity is preserved. Expected results are specialised methods for the extraction, the selection, and the adaptation of axioms to integrate.

Incremental ontology evolution

There are several actively maintained knowledge graphs such as DBpedia, Yago, or Wikidata that evolve their data regularly. While changes mainly happen at the level of instances (creation, property value change, etc.), they affect more rarely the ontologies. This can be explained because changes at the level of the ontologies are more critical and complex to implement than those at the instance level. However ontologies also have to be adapted to better reflect the reality, if ontologies do not evolve, the knowledge will become obsolete and die. It therefore seems important to us to be able to anticipate the evolution of ontologies based on the dynamics of the data.

Ontology evolution from instances, also called data-driven change discovery, mostly rely on ontology learning methods (Lehmann and Hitzler, 2010; Mädche, 2002). These methods have not been especially designed to make ontologies evolve and thus they do not take existing ontologies into account. Furthermore, learning methods often require a large number of instances to statically assess the quality of extracted patterns. Only one work has focused specifically on learning ontology evolution from instances (Saïs, Pruski, et al., 2017).

Our approach to handle the problem will take advantage of both symbolic data-mining methods, ontologies, and adaptation operators. The use of symbolic data mining ap-

proaches, such as those based on FCA or RCA, is suitable because they are not restricted to frequent patterns. The role of the ontology to evolve will also be central because it will guide the extraction and can guess the quality of extracted axiom thanks to consistency. To design adaptation operators, several directions have to be investigated and combined both at the experimental and theoretical levels. They may be extracted from ontology versioning logs or inspired by those studied in cultural knowledge evolution (Euzenat, 2017; Bourahla, Atencia, et al., 2021) for instance. Their properties can follow some of the principles of belief revision such as AGM postulates (Alchourrón et al., 1985), and studied through dynamic epistemic logics (van den Berg et al., 2021).

We expect that combining such extraction methods and ontological knowledge will allow to learn accurately from relatively little amount of data. Another achievement will to be able to enhance the global quality of knowledge graphs. Indeed, we expect that axioms discovered by such methods will help to detect exceptions in data that represent errors to be corrected.

Measuring knowledge co-evolution

Measuring the evolution of knowledge is essential in many aspects. It can provide insights into the evolutionary dynamics of a knowledge graph and answer many questions such as: Are the graph data up to date? Is the knowledge of the ontology stable, does the graph become more similar to other knowledge structures over time?, etc. Even if the first two questions can be partially answered by analyzing the evolution of a single graph, their joint analysis would allow to obtain more precise answers.

In the literature, measures of knowledge evolution are only concerned with changes in the same structure. For instance, (Pernisch et al., 2021) considered measuring the impacts of changes between versions of ontologies. In (Orme et al., 2007), authors studied if the evolution of standard complexity and cohesion metrics can capture changes and how they can reflect the stability of an ontology over time. None of these works addressed the evolution of ontologies globally in relation with the other knowledge structures.

Our approach for measuring global knowledge evolution takes inspiration from phylogenetics. Indeed, knowledge representations evolve over time like biological organisms and representations can share common ancestors like DBPedia, Yago, Google KG which both use informal knowledge from Wikipedia. Phylogenetics proposes several tools for analysing the evolution. For instance, phylogenetic inference methods allow to build phylogenetic trees from data traits. Several different trees can be built, each of them representing a particular hypothesis about the evolution of genes or species. By conjointly exploiting our expertise on ontology measures (see Chapter 2), alignments, and classification methods, we could test different hypothesises of the evolution of knowledge. Considered traits can be, for instance, textual annotations, common instances, shared axiom or structural similarity. From such structures, we will define original measures quantifying the diversity (Tucker et al., 2016) or the stability of knowledge over time. We expect to apply these measures in the context of cultural knowledge evolution for controlling these aspects during simulations.

Appendix A

The OntoSim library

This work on distances between ontologies has yield the OntoSim Library [1]. We have designed this library in order to be extensible and we have implemented all the measures presented in (David and Euzenat, 2008a) and (David, Euzenat, and Sváb-Zamazal, 2010). This software is written in Java, is quite independent of the ontology API (JENA or OWL API) and is connected to the Alignment API.

OntoSim is based on a minimal but generic `Measure` interface that a concrete class has to implement:

```
public interface Measure<O> {
    static enum TYPES {similarity, dissimilarity, distance, other};

    public TYPES getMType();
    public double getMeasureValue( O o1, O o2);
    public double getSim( O o1, O o2);
    public double getDissim( O o1, O o2);
}
```

OntoSim is provided with two vector space measures (boolean and $TF \cdot IDF$), four concept-based measures and four alignment space measures. In addition, the framework can embeds external similarity libraries which can be combined with our owns.

It also provides various tools for creating new concept-based measures that combine values from a matrix. This is materialized by the `SetMeasure` interface that is parametrized by a local `Measure`, and `Extractor`, and an `AggregationScheme`. In particular, we have implemented various extractors such as basic thresholding, max, min, max-min, stable marriage, Hungarian algorithm and also different kinds of aggregation schemes such as generalized mean or weighted sum.

This software has been designed to facilitate its extension and its integration by other tools, such as matchers, through its API. It is freely available under LGPL 2.1.

[1] `https://gitlab.inria.fr/moex/ontosim/`

Appendix B

Linkex: Link key extraction tool

All our work on discovery algorithms and link key quality measures, except the RCA algorithm for dependent link keys extraction, are implemented in the Linkex tool[1], written in Java.

This tool takes as input two RDF datasets given as files in Turtle, RDF/XML or NTriple format and outputs the extracted link key candidatesin the Alignement API EDOAL format.

The link key extraction procedure consists of the following steps:

1. The indexation and preprocessing of RDF datsets;
2. The construction of the set of descriptions that associates for each pair of instances, the sets of pair of properties for which the instances share at least one value or all values;
3. The computation of the lattice of link key candidates (with a modified version of the addIntent algorithm (Merwe et al., 2004));
4. The quality evaluation of candidates;
5. The rendering of candidates.

During the indexation phase, values are normalized in order to reduce their heterogeneity. The transformation consists in (1) removing diacritics, (2) tokenizing (using sequence of non digit or letter as separator), (3) sorting the resulting bag of token. Linkex allows to take into account composition of properties (and inverse properties). Properties can also be filtered using a minimal support and/or a discriminability threshold.

The resulting link key candidates can be rendered in EDOAL[2], but also in a special text format where each link key is associated with several quality measure values or as a GraphViz dot file representing the lattice.

[1] `https://gitlab.inria.fr/moex/linkex`

[2] `https://moex.gitlabpages.inria.fr/alignapi/edoal.html`

Bibliography

Abbas, Nacira, Jérôme David, and Amedeo Napoli (2020). "Discovery of link keys in RDF data based on pattern structures: preliminary steps". en. In: *Proc. 15th International conference on Concept Lattices and their Applications (CLA), Tallinn (EE)*. Ed. by Francisco José Valverde-Albacete and Martin Trnecka, pp. 235–246. URL: http://ceur-ws.org/Vol-2668/paper18.pdf.

Achichi, Manel, Michelle Cheatham, Zlatan Dragisic, Jérôme Euzenat, Daniel Faria, Alfio Ferrara, Giorgos Flouris, Irini Fundulaki, Ian Harrow, Valentina Ivanova, Ernesto Jiménez-Ruiz, Elena Kuss, Patrick Lambrix, Henrik Leopold, Huanyu Li, Christian Meilicke, Stefano Montanelli, Catia Pesquita, Tzanina Saveta, Pavel Shvaiko, Andrea Splendiani, Heiner Stuckenschmidt, Konstantin Todorov, Cássia Trojahn dos Santos, and Ondrej Zamazal (2016). "Results of the Ontology Alignment Evaluation Initiative 2016". en. In: *Proc. 11th ISWC workshop on ontology matching (OM), Kobe (JP)*. Ed. by Pavel Shvaiko, Jérôme Euzenat, Ernesto Jiménez-Ruiz, Michelle Cheatham, Oktie Hassanzadeh, and Ryutaro Ichise, pp. 73–129. URL: http://oaei.ontologymatching.org/2016/results/oaei2016.pdf.

Achichi, Manel, Mohamed Ben Ellefi, Danai Symeonidou, and Konstantin Todorov (2016). "Automatic key selection for data linking". In: *Proc. 20th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2016), Bologna, Italy*. Ed. by Eva Blomqvist, Paolo Ciancarini, Francesco Poggi, and Fabio Vitali. Vol. 10024. Lecture Notes in Computer Science. doi: 10.1007/978-3-319-49004-5_1. Springer, pp. 3–18.

Alchourrón, Carlos E., Peter Gärdenfors, and David Makinson (1985). "On the Logic of Theory Change: Partial Meet Contraction and Revision Functions". In: *Journal of Symbolic Logic* 50.2, pp. 510–530. DOI: 10.2307/2274239.

Atencia, Manuel, Michel Chein, Madalina Croitoru, Jérôme David, Michel Leclère, Nathalie Pernelle, Fatiha Saïs, François Scharffe, and Danai Symeonidou (2014). "Defining key semantics for the RDF datasets: experiments and evaluations". en. In: *Proc. 21st International Conference on Conceptual Structures (ICCS), Iasi (RO)*. Vol. 8577. Lecture notes in artificial intelligence, pp. 65–78. URL: https://exmo.inria.fr/files/publications/atencia2014c.pdf.

Atencia, Manuel, Jérôme David, and Jérôme Euzenat (2014a). "Data interlinking through robust linkkey extraction". en. In: *Proc. 21st european conference on artificial intelligence (ECAI), Praha (CZ)*. Ed. by Torsten Schaub, Gerhard Friedrich, and Barry O'Sullivan. Amsterdam (NL): IOS press, pp. 15–20. URL: https://exmo.inria.fr/files/publications/atencia2014b.pdf.

— (2014b). "What can FCA do for database linkkey extraction?" en. In: *Proc. 3rd ECAI workshop on What can FCA do for Artificial Intelligence? (FCA4AI), Praha (CZ)*, pp. 85–92. URL: http://ceur-ws.org/Vol-1257/paper10.pdf.

Atencia, Manuel, Jérôme David, and Jérôme Euzenat (2019). "Several link keys are better than one, or extracting disjunctions of link key candidates". en. In: *Proc. 10th ACM international conference on knowledge capture (K-Cap), Marina del Rey (CA US)*, pp. 61–68. URL: https://moex.inria.fr/files/papers/atencia2019c.pdf.

— (2021). "On the relation between keys and link keys for data interlinking". en. In: *Semantic web journal* 12.4, pp. 547–567. URL: https://content.iospress.com/articles/semantic-web/sw200414.

Atencia, Manuel, Jérôme David, Jérôme Euzenat, Amedeo Napoli, and Jérémy Vizzini (2020). "Link key candidate extraction with relational concept analysis". en. In: *Discrete applied mathematics* 273, pp. 2–20. URL: https://moex.inria.fr/files/papers/atencia2019z.pdf.

Atencia, Manuel, Jérôme David, and François Scharffe (2012). "Keys and pseudo-keys detection for web datasets cleansing and interlinking". en. In: *Proc. 18th international conference on knowledge engineering and knowledge management (EKAW), Galway (IE)*, pp. 144–153. URL: https://exmo.inria.fr/files/publications/atencia2012b.pdf.

Baixeries, Jaume, Victor Codocedo, Mehdi Kaytoue, and Amedeo Napoli (2018). "Characterizing Approximate-Matching Dependencies in Formal Concept Analysis with Pattern Structures". In: *Discrete Applied Mathematics* 249, pp. 18–27. DOI: 10.1016/j.dam.2018.03.073. URL: https://hal.inria.fr/hal-01673441.

Bizer, Christian, Julius Volz, Georgi Kobilarov, and Martin Gaedke (2009). "Silk - A Link Discovery Framework for the Web of Data". In: *Linked Data on the Web Workshop (LDOW09), Workshop at 18th International World Wide Web Conference (WWW09)*. Madrid (ES).

Bourahla, Yasser, Manuel Atencia, and Jérôme Euzenat (2021). "Knowledge improvement and diversity under interaction-driven adaptation of learned ontologies". en. In: *Proc. 20th ACM international conference on Autonomous Agents and Multi-Agent Systems (AAMAS), London (UK)*. Ed. by Ulle Endriss, Ann Nowé, Frank Dignum, and Alessio Lomuscio, pp. 242–250. URL: http://www.ifaamas.org/Proceedings/aamas2021/pdfs/p242.pdf.

Bourahla, Yasser, Jérôme David, Jérôme Euzenat, and Meryem Naciri (2022). "Measuring and controlling knowledge diversity". en. In: *Proc. 1st JOWO workshop on formal models of knowledge diversity (FMKD), Jönköping (SE)*. URL: https://moex.inria.fr/files/papers/bourahla2022c.pdf.

Brickley, Dan and Ramanathan V. Guha (2014). *RDF Schema 1.1*. Tech. rep. W3C. URL: https://www.w3.org/TR/rdf-schema/.

Buitelaar, Paul, Philipp Cimiano, and Bernardo Magnini, eds. (2005). *Ontology Learning from Text: Methods, Evaluation and Applications*. Vol. 123. Frontiers in Artificial Intelligence and Applications. IOS Press.

Chen, Jiaoyan, Pan Hu, Ernesto Jiménez-Ruiz, Ole Magnus Holter, Denvar Antonyrajah, and Ian Horrocks (2021). "OWL2Vec*: embedding of OWL ontologies". In: *Machine Learning* 110.7, pp. 1813–1845. URL: https://doi.org/10.1007/s10994-021-05997-6.

Christen, Peter (2012). *Data Matching—Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer, Heidelberg (DE).

d'Aquin, Mathieu (2009). "Formally Measuring Agreement and Disagreement in Ontologies". In: *Proc. of the Fifth International Conference on Knowledge Capture*. K-CAP

'09. Redondo Beach, California, USA: ACM, pp. 145–152. ISBN: 978-1-60558-658-8. DOI: 10.1145/1597735.1597761. URL: http://doi.acm.org/10.1145/1597735.1597761.

David, Jérôme and Jérôme Euzenat (2008a). "Comparison between ontology distances (preliminary results)". In: *Proc. 7th international semantic web conference (ISWC), Karlsruhe (DE)*. Vol. 5318. Lecture notes in computer science, pp. 245–260. URL: https://exmo.inria.fr/files/publications/david2008a.pdf.

— (2008b). "On fixing semantic alignment evaluation measures". en. In: *Proc. 3rd ISWC workshop on ontology matching (OM), Karlsruhe (DE)*. Ed. by Pavel Shvaiko, Jérôme Euzenat, Fausto Giunchiglia, and Heiner Stuckenschmidt, pp. 25–36. URL: https://exmo.inria.fr/files/publications/david2008b.pdf.

David, Jérôme, Jérôme Euzenat, Pierre Genevès, and Nabil Layaïda (2018). "Evaluation of query transformations without data". en. In: *Proc. WWW workshop on Reasoning on Data (RoD), Lyon (FR)*. ACM Press, pp. 1599–1602. URL: https://moex.inria.fr/files/papers/david2018a.pdf.

David, Jérôme, Jérôme Euzenat, and Ondrej Sváb-Zamazal (2010). "Ontology similarity in the alignment space". en. In: *Proc. 9th international semantic web conference (ISWC), Shanghai (CN)*, pp. 129–144. URL: https://exmo.inria.fr/files/publications/david2010b.pdf.

David, Jérôme, Fabrice Guillet, and Henri Briand (2007). "Association Rule Ontology Matching Approach". en. In: *International journal of semantic web and information systems* 3.2, pp. 27–49.

Ehrig, Marc and Jérôme Euzenat (2005). "Relaxed Precision and Recall for Ontology Matching". en. In: *Proc. K-CAP Workshop on Integrating Ontologies*. Banff (CA), pp. 25–32. URL: http://ceur-ws.org/Vol-156/paper5.pdf.

Ehrig, Marc, Peter Haase, Mark Hefke, and Nenad Stojanovic (2005). "Similarity for Ontologies – A Comprehensive Framework". In: *Proc. 13th European Conference on Information Systems, Information Systems in a Rapidly Changing Economy (ECIS), Regensburg (DE)*.

Euzenat, Jérôme (2003). "Towards composing and benchmarking ontology alignments". en. In: *Proc. ISWC Workshop on Semantic Integration*. Sanibel Island (FL US), pp. 165–166.

— (2007). "Semantic precision and recall for ontology alignment evaluation". In: *Proc. 20th International Joint Conference on Artificial Intelligence (IJCAI)*. Hyderabad (IN), pp. 248–253.

— (2008). "Algebras of ontology alignment relations". en. In: *Proc. 7th conference on international semantic web conference (ISWC), Karlsruhe (DE)*. Vol. 5318. Lecture notes in computer science, pp. 387–402. URL: http://www.springerlink.com/index/DY8Y9F31A9GT9762.

— (2017). "Interaction-based ontology alignment repair with expansion and relaxation". en. In: *Proc. 26th International Joint Conference on Artificial Intelligence (IJCAI), Melbourne (VIC AU)*. Menlo Park (CA US): AAAI Press, pp. 185–191. URL: http://static.ijcai.org/proceedings-2017/0027.pdf.

Euzenat, Jérôme and Pavel Shvaiko (2007). *Ontology matching*. Heidelberg (DE): Springer.

— (2013). *Ontology matching*. en. 2nd. Heidelberg (DE): Springer-Verlag. 520 pp.

Euzenat, Jérôme and Petko Valtchev (2004). "Similarity-based ontology alignment in OWL-Lite". In: *Proc. 16th European Conference on Artificial Intelligence (ECAI)*. Valencia (ES), pp. 333–337.

Farah, Houssameddine, Danai Symeonidou, and Konstantin Todorov (2017). "KeyRanker: Automatic RDF key ranking for data linking". In: *Proc. the Knowledge Capture Conference (K-CAP)*. Ed. by Óscar Corcho, Krzysztof Janowicz, Giuseppe Rizzo, Ilaria Tiddi, and Daniel Garijo. doi: 10.1145/3148011.3148023. New York, NY, USA: ACM, 7:1–7:8.

Ferrara, Alfio, Andriy Nikolov, and François Scharffe (2011). "Data Linking for the Semantic Web". In: *International Journal of Semantic Web and Information Systems* 7.3, pp. 46–76.

Ferré, Sébastien (2021). "Application of concepts of neighbours to knowledge graph completion". In: *CODATA Data Science Journal* 4.1, pp. 1–28. DOI: 10.3233/DS-200030. URL: https://hal.inria.fr/hal-03531781.

Flouris, Giorgos, Dimitris Manakanatas, Haridimos Kondylakis, Dimitris Plexousakis, and Grigoris Antoniou (2008). "Ontology Change: Classification and Survey". In: *Knowledge Engineering Review* 23.2, pp. 117–152. ISSN: 0269-8889. DOI: 10.1017/S0269888908001367. URL: https://doi.org/10.1017/S0269888908001367.

Ganter, Bernhard and Sergei O. Kuznetsov (2001). "Pattern Structures and Their Projections". In: *Proc. 9th International Conference on Conceptual Structures (ICCS 2001)*. LNCS 2120. Springer, pp. 129–142.

Group, W3C OWL Working (2012). *OWL 2 Web Ontology Language Document Overview (Second Edition)*. Tech. rep. W3C. URL: https://www.w3.org/TR/owl2-overview/.

Heath, Tom and Christian Bizer (2011). *Linked Data : Evolving the Web into a global data space*. doi: 10.2200/S00334ED1V01Y201102WBE001. Morgan and Claypool.

Hogan, Aidan, Antoine Zimmermann, Jürgen Umbrich, Axel Polleres, and Stefan Decker (2012). "Scalable and distributed methods for entity matching, consolidation and disambiguation over linked data corpora". In: *Journal of Web Semantics* 10, pp. 76–110.

Hong, Lu and Scott Page (2004). "Groups of diverse problem solvers can outperform groups of high-ability problem solvers". In: *Proceedings of the national academy of sciences* 101.46, pp. 16385–16389.

Horridge, Matthew and Peter Patel-Schneider (2012). *OWL 2 Web Ontology Language. Manchester Syntax (Second Edition)*. URL: http://www.w3.org/TR/owl2-manchester-syntax/ (visited on 03/15/2015).

Hu, Bo, Yannis Kalfoglou, Harith Alani, David Dupplaw, Paul Lewis, and Nigel Shadbolt (2006). "Semantic Metrics". In: *Proc. 15th International Conference on Knowledge Engineering and Knowledge Management (EKAW)*. Vol. 4248. Lecture notes in computer science. Praha (CZ), pp. 166–181.

Huhtala, Ykä, Juha Kärkkäinen, Pasi Porkka, and Hannu Toivonen (1999). "TANE: An Efficient Algorithm for Discovering Functional and Approximate Dependencies". In: *Comput. J.* 42.2, pp. 100–111.

Inants, Armen (2016). "Qualitative calculi with heterogeneous universes. (Calculs qualitatifs avec des univers hétérogènes)". PhD thesis. Grenoble Alpes University, France. URL: https://tel.archives-ouvertes.fr/tel-01366032.

Isele, Robert and Christian Bizer (2013). "Active learning of expressive linkage rules using genetic programming". In: *Journal of web semantics* 23, pp. 2–15.

Isele, Robert, Anja Jentzsch, and Christian Bizer (2011). "Efficient multidimensional blocking for link discovery without losing recall". In: *Proc. the 14th International Workshop on the Web and Databases 201 (WebDB 2011)*. Ed. by Amélie Marian and Vasilis Vassalos. Athens, Greece.

Jiménez-Ruiz, Ernesto, Bernardo Cuenca Grau, Ian Horrocks, and Rafael Berlanga (2011). "Logic-based assessment of the compatibility of UMLS ontology sources". In: *Journal of Biomedical Semantics* 2.1, S2. DOI: 10.1186/2041-1480-2-S1-S2. URL: https://doi.org/10.1186/2041-1480-2-S1-S2.

Kuhn, Harold W. (Mar. 1955). "The Hungarian Method for the Assignment Problem". In: *Naval Research Logistics Quarterly* 2.1–2, pp. 83–97. DOI: 10.1002/nav.3800020109.

Lakhal, Lotfi and Gerd Stumme (2005). "Efficient Mining of Association Rules Based on Formal Concept Analysis". In: *Formal Concept Analysis: Foundations and Applications*. Ed. by Bernhard Ganter, Gerd Stumme, and Rudolf Wille, pp. 180–195. ISBN: 978-3-540-31881-1. DOI: 10.1007/11528784_10.

Lehmann, Jens and Pascal Hitzler (2010). "Concept Learning in Description Logics Using Refinement Operators". In: *Machine Learning journal* 78.1-2, pp. 203–250.

Leinster, Tom (2021). *Entropy and diversity: the axiomatic approach*. Cambridge (UK): Cambridge university press. ISBN: 9781108965576. URL: https://arxiv.org/pdf/2012.02113.pdf.

Lesnikova, Tatiana (2016). "RDF data interlinking: evaluation of cross-lingual methods". en. Thèse d'informatique. Grenoble (FR): Université de Grenoble. URL: https://exmo.inria.fr/files/thesis/thesis-lesnikova.pdf.

Locoro, Angela, Jérôme David, and Jérôme Euzenat (2014). "Context-based matching: design of a flexible framework and experiment". en. In: *Journal on data semantics* 3.1, pp. 25–46. URL: https://exmo.inria.fr/files/publications/locoro2014a.pdf.

Mädche, Alexander (2002). *Ontology Learning for the Semantic Web*. Kluwer Academic Publishers.

Mädche, Alexander and Steffen Staab (2002). "Measuring Similarity between Ontologies". In: *Proc. 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW)*. Vol. 2473. Lecture notes in computer science. Siguenza (ES), pp. 251–263.

Mannila, Heikki and Kari-Jouko Raiha (1994). "Algorithms for inferring functional-dependencies from relations". In: *Data & Knowledge Engineering* 12, pp. 83–99.

Meilicke, Christian (2011). "Alignment Incoherence in Ontology Matching". English. PhD thesis. Mannheim (DE): Universität Mannheim. URL: https://madoc.bib.uni-mannheim.de/29351/.

Meilicke, Christian, Melisachew Wudage Chekol, Daniel Ruffinelli, and Heiner Stuckenschmidt (July 2019). "Anytime Bottom-Up Rule Learning for Knowledge Graph Completion". In: *Proc. 28th international Joint Conference on Artificial Intelligence (IJCAI-19)*, pp. 3137–3143. DOI: 10.24963/ijcai.2019/435. URL: https://doi.org/10.24963/ijcai.2019/435.

Meilicke, Christian and Heiner Stuckenschmidt (2008). "Incoherence as a Basis for Measuring the Quality of Ontology Mappings". In: *Proc. the 3rd International Conference on Ontology Matching - Volume 431*. OM'08. Karlsruhe, Germany: CEUR-WS.org, pp. 1–12.

Merwe, Dean van der, Sergei Obiedkov, and Derrick Kourie (2004). "AddIntent: A New Incremental Algorithm for Constructing Concept Lattices". In: *Proc. 2nd International Conference on Formal Concept Analysis (ICFCA 2004)*. Ed. by Peter Eklund. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 372–385.

Miles, Alistair and Sean Bechhofer (2009). *SKOS Simple Knowledge Organization System Reference*. Tech. rep. W3C. URL: https://www.w3.org/TR/skos-reference/.

Miller, George (1995). "WordNet: A lexical database for English". In: *Communications of the ACM* 38.11, pp. 39–41.

Nentwig, Markus, Michael Hartung, Axel-Cyrille Ngonga Ngomo, and Erhard Rahm (2017a). "A survey of current Link Discovery frameworks". In: *Semantic Web* 8.3, pp. 419–436.

— (2017b). "A survey of current link discovery frameworks". In: *Semantic Web* 8.3. doi: 10.3233/SW-150210, pp. 419–436.

Ngonga Ngomo, Axel-Cyrille (2014). "HELIOS – Execution Optimization for Link Discovery". In: *proc. 13th International Semantic Web Conference (ISWC 2014)*. Cham: Springer, pp. 17–32.

Ngonga Ngomo, Axel-Cyrille and Sören Auer (2011). "LIMES: A Time-Efficient Approach for Large-Scale Link Discovery on the Web of Data". In: *Proc. 22nd International Joint Conference on Artificial Intelligence (IJCAI)*. Barcelona (ES), pp. 2312–2317.

Ngonga Ngomo, Axel-Cyrille and Klaus Lyko (2012). "EAGLE: Efficient Active Learning of Link Specifications Using Genetic Programming". In: *Proc. 9th Extended Semantic Web Conference (ESWC 2012)*. Vol. 7295. Lecture Notes in Computer Science. Heraklion (GR): Springer, pp. 149–163.

Ngonga Ngomo, Axel-Cyrille, Klaus Lyko, and Victor Christen (2013). "COALA - Correlation-Aware Active Learning of Link Specifications". In: *Proc. 10th Extended Semantic Web Conference (ESWC 2013)*. Vol. 7882. Lecture Notes in Computer Science. Montpellier (FR): Springer, pp. 442–456.

Nikolov, Andriy, Mathieu d'Aquin, and Enrico Motta (2012). "Unsupervised Learning of Link Discovery Configuration". In: *Proc. of the 9th Extended Semantic Web Conference (ESWC 12)*. Ed. by Elena Simperl, Philipp Cimiano, Axel Polleres, Oscar Corcho, and Valentina Presutti. Springer, pp. 119–133.

Nikolov, Andriy, Alfio Ferrara, and François Scharffe (July 2011). "Data Linking for the Semantic Web". In: *Int. J. Semant. Web Inf. Syst.* 7.3, pp. 46–76. ISSN: 1552-6283. DOI: 10.4018/jswis.2011070103. URL: https://doi.org/10.4018/jswis.2011070103.

Nikolov, Andriy, Victoria Uren, Enrico Motta, and Anne de Roeck (2008). "Handling instance coreferencing in the KnoFuss architecture". In: *Proc. of the workshop: Identity and Reference on the Semantic Web at 5th European Semantic Web Conference (ESWC 2008)*.

Noble, Diego, Marcelo Prates, Daniel Bossle, and Luís Lamb (2015). "Collaboration in Social Problem-Solving: When Diversity Trumps Network Efficiency". In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. AAAI'15. Austin, Texas: AAAI Press, pp. 1277–1283.

Orme, Anthony M., Haining Yao, and Letha H. Etzkorn (2007). "Indicating ontology data quality, stability, and completeness throughout ontology evolution". In: *Journal of Software Maintenance and Evolution: Research and Practice* 19.1, pp. 49–75. DOI: https://doi.org/10.1002/smr.341. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/smr.341.

Paulheim, Heiko (2017). "Knowledge graph refinement: A survey of approaches and evaluation methods". In: *Semantic web* 8.3, pp. 489–508.

Pernisch, Romana, Daniele Dell'Aglio, and Abraham Bernstein (2021). "Beware of the Hierarchy — An Analysis of Ontology Evolution and the Materialisation Impact for Biomedical Ontologies". In: *Web Semantics* 70.C. ISSN: 1570-8268. DOI: 10.1016/j.websem.2021.100658. URL: https://doi.org/10.1016/j.websem.2021.100658.

Pirrò, Giuseppe (2019). "Semantic Similarity Functions and Measures". In: *Encyclopedia of Bioinformatics and Computational Biology*. Ed. by Shoba Ranganathan, Michael Gribskov, Kenta Nakai, and Christian Schönbach. Oxford: Academic Press, pp. 877–888. ISBN: 9782-12-811432-2. DOI: https://doi.org/10.1016/B978-0-12-809633-8.20402-0. URL: https://www.sciencedirect.com/science/article/pii/B9780128096338204020.

Pirrò, Giuseppe and Jérôme Euzenat (Nov. 2010). "A feature and information theoretic framework for semantic similarity and relatedness". In: *Proc. 9th international semantic web conference (ISWC)*. Ed. by Peter Patel-Schneider, Yue Pan, Pascal Hitzler, Peter Mika, Lei Zhang, Jeff Pan, Ian Horrocks, and Birte Glimm. Vol. 6496. Lecture notes in computer science. pirro2010b. Shanghai, China: Springer Verlag, pp. 615–630. DOI: 10.1007/978-3-642-17746-0_39. URL: https://hal.inria.fr/hal-00793283.

Resnik, Philipp (1995). "Using information content to evaluate semantic similarity in a taxonomy". In: *Proc. 14th International Joint Conference on Artificial Intelligence (IJCAI)*. Montréal (CA), pp. 448–453.

Rouane-Hacene, Mohamed, Marianne Huchard, Amedeo Napoli, and Petko Valtchev (2013). "Relational Concept Analysis: mining concept lattices from multi-relational data". In: *Annals of Mathematics and Artificial Intelligence* 67.1, pp. 81–108.

Saïs, Fatiha, Nathalie Pernelle, and Marie-Christine Rousset (2007). "L2R: A Logical Method for Reference Reconciliation". In: *Proc. 22$^{nd]}$ AAAI Conference on Artificial Intelligence*. Vancouver (CA), pp. 329–334.

— (2009). "Combining a Logical and a Numerical Method for Data Reconciliation". In: *Journal on Data Semantics* 12.12, pp. 66–94. URL: https://hal.inria.fr/inria-00433007.

Saïs, Fatiha, Cédric Pruski, and Marcos Da Silveira (2017). "Inferring the Evolution of Ontology Axioms from RDF Data Dynamics". In: *proc. of 9th Knowledge Capture Conference (K-CAP)*. K-CAP 2017. Austin, TX, USA: ACM. ISBN: 9781450355537. DOI: 10.1145/3148011.3154472. URL: https://doi.org/10.1145/3148011.3154472.

Scharffe, François and Jérôme Euzenat (2011). "MeLinDa: an interlinking framework for the web of data". In: *CoRR* abs/1107.4502.

Scharffe, François, Yanbin Liu, and Chunguang Zhou (2009). "RDF-AI: an Architecture for RDF Datasets Matching, Fusion and Interlink". In: *Proc of the Workshop on Identity and Reference in Knowledge Representation, IJCAI 2009*.

Seco, Nuno, Tony Veale, and Jer Hayes (2004). "An Intrinsic Information Content Metric for Semantic Similarity in WordNet". In: *Proc. 16th European Conference on Artificial Intelligence (ECAI)*. ECAI'04. Valencia, Spain: IOS Press, pp. 1089–1090. ISBN: 9781586034528.

Sherif, Mohamed Ahmed, Axel-Cyrille Ngonga Ngomo, and Jens Lehmann (2017). "Wombat – A Generalization Approach for Automatic Link Discovery". In: *proc. 14th International Semantic Web Conference (ISWC)*. Ed. by Eva Blomqvist, Diana Maynard, Aldo Gangemi, Rinke Hoekstra, Pascal Hitzler, and Olaf Hartig. Springer, pp. 103–119. ISBN: 978-3-319-58068-5.

Sismanis, Yannis, Paul Brown, Peter J. Haas, and Berthold Reinwald (2006). "GORDIAN: efficient and scalable discovery of composite keys". In: *Proc. of the 32nd international conference on Very large data bases*. VLDB '06. Seoul, Korea: VLDB Endow-

ment, pp. 691–702. URL: http://dl.acm.org/citation.cfm?id=1182635.1164187.

Solimando, Alessandro, Ernesto Jiménez-Ruiz, and Giovanna Guerrini (2014). "Detecting and Correcting Conservativity Principle Violations in Ontology-to-Ontology Mappings". In: *The Semantic Web – ISWC 2014*. Ed. by Peter Mika, Tania Tudorache, Abraham Bernstein, Chris Welty, Craig Knoblock, Denny Vrandečić, Paul Groth, Natasha Noy, Krzysztof Janowicz, and Carole Goble. Cham: Springer International Publishing, pp. 1–16. ISBN: 978-3-319-11915-1.

— (2017). "Minimizing conservativity violations in ontology alignments: algorithms and evaluation". In: *Knowledge and Information Systems* 51.3, pp. 775–819. ISSN: 0219-3116. DOI: 10.1007/s10115-016-0983-3. URL: https://doi.org/10.1007/s10115-016-0983-3.

Solimando, Alessandro, Ernesto Jiménez-Ruiz, and Christoph Pinkel (2014). "Evaluating Ontology Alignment Systems in Query Answering Tasks". In: *Proc. 2014 International Semantic Web Conference - Posters and Demonstrations Track - Volume 1272*. ISWC-PD'14. Riva del Garda, Italy: CEUR-WS.org, pp. 301–304.

Song, Dezhao and Jeff Heflin (2011). "Automatically Generating Data Linkages Using a Domain-Independent Candidate Selection Approach". In: *Proc of the 10th International Semantic Web Conference (ISWC11)*, pp. 649–664.

Song, Shaoxu and Lei Chen (2009). "Discovering Matching Dependencies". In: *Proceedings of the 18th ACM Conference on Information and Knowledge Management*. CIKM '09. Hong Kong, China: Association for Computing Machinery, pp. 1421–1424. ISBN: 9781605585123. DOI: 10.1145/1645953.1646135. URL: https://doi.org/10.1145/1645953.1646135.

Soru, Tommaso, Edgard Marx, and Axel-Cyrille Ngonga Ngomo (2015). "ROCKER: A Refinement Operator for Key Discovery". In: *Proc. 24th International Conference on World Wide Web (WWW)*. Florence (IT): International World Wide Web Conferences Steering Committee, pp. 1025–1033.

Stirling, Andy (2007). "A general framework for analysing diversity in science, technology and society". In: *Journal of the royal society—Interface* 4, pp. 707–719.

Stuckenschmidt, Heiner, Luciano Serafini, and Holger Wache (2006). "Reasoning about Ontology Mappings". In: *Proc. ECAI 2006 workshop on contextual representation and reasoning*. Riva del Garda, Italy.

Šváb, Ondřej, Vojtěch Svátek, Petr Berka, Dušan Rak, and Petr Tomášek (2005). "OntoFarm: Towards an Experimental Collection of Parallel Ontologies". In: *Proc. 4th ISWC poster session, Galway (IE)*.

Symeonidou, Danai, Vincent Armant, Nathalie Pernelle, and Fatiha Saïs (2014). "SAKey: Scalable Almost Key Discovery in RDF Data". In: *Proc. 13th International Semantic Web Conference (ISWC 2014)*. Ed. by Peter Mika, Tania Tudorache, Abraham Bernstein, Chris Welty, Craig A. Knoblock, Denny Vrandecic, Paul Groth, Natasha F. Noy, Krzysztof Janowicz, and Carole A. Goble. Vol. 8796. Lecture Notes in Computer Science. doi: 10.1007/978-3-319-11964-9_3. Springer, pp. 33–49.

Symeonidou, Danai, Nathalie Pernelle, and Fatiha Saïs (2011). "KD2R: A Key Discovery Method for Semantic Reference Reconciliation". In: *Proc. 7th International IFIP Workshop on Semantic Web and Web Semantics*. Vol. 7046. OTM Workshops. Hersonissos (GR): Springer-Verlag.

Thiéblin, Élodie, Ollivier Haemmerlé, Nathalie Hernandez, and Cassia Trojahn (2020). "Survey on Complex Ontology Matching". In: *Semantic Web* 11.4, pp. 689–727. ISSN:

1570-0844. DOI: 10.3233/SW-190366. URL: https://doi.org/10.3233/SW-190366.

Thiéblin, Élodie, Ollivier Haemmerlé, and Cássia Trojahn (Jan. 2021). "Automatic Evaluation of Complex Alignments: An Instance-Based Approach". In: *Semantic Web* 12.5, pp. 767–787. ISSN: 1570-0844. DOI: 10.3233/SW-210437.

Tucker, C, M Cadotte, S Carvalho, T Davies, S Ferrier, S Fritz, R Grenyer, M Helmus, L Jin, A Mooers, S Pavoine, O Purschke, D Redding, D Rosauer, M Winter, and F Mazel (2016). "A guide to phylogenetic metrics for conservation, community ecology and macroecology." In: *Biological Reviews of the Cambridge Philosophical Society* 2.92, pp. 698–715.

Valtchev, Petko (1999). "Construction automatique de taxonomies pour l'aide à la représentation de connaissances par objets". Thèse d'informatique. Grenoble (FR): Université Grenoble 1.

van den Berg, Line, Manuel Atencia, and Jérôme Euzenat (2021). "A logical model for the ontology alignment repair game". en. In: *Autonomous agents and multi-agent systems* 35.2, p. 32. URL: https://moex.inria.fr/files/papers/vandenberg2021a.pdf.

Volz, Julius, Christian Bizer, Martin Gaedke, and Georgi Kobilarov (2009). "Discovering and Maintaining Links on the Web of Data". In: *proc. 8th International Semantic Web Conference (ISWC 2009)*. Ed. by Abraham Bernstein, David R. Karger, Tom Heath, Lee Feigenbaum, Diana Maynard, Enrico Motta, and Krishnaprasad Thirunarayan. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 650–665.

Vrandečić, Denny and York Sure (2007). "How to Design Better Ontology Metrics". In: *Proc. 4th European Semantic Web Conference, Innsbruck (AT)*. Vol. 4519. Lecture Notes in Computer Science, pp. 311–325.

Wang, Quan, Zhendong Mao, Bin Wang, and Li Guo (2017). "Knowledge Graph Embedding: A Survey of Approaches and Applications". In: *IEEE Transactions on Knowledge and Data Engineering* 29.12, pp. 2724–2743. DOI: 10.1109/TKDE.2017.2754499.

Zablith, Fouad, Grigoris Antoniou, Mathieu d'Aquin, Giorgos Flouris, Haridimos Kondylakis, Enrico Motta, Dimitris Plexousakis, and Marta Sabou (2016). "Ontology evolution: a process-centric survey". In: *The Knowledge Engineering Review* 30.1, pp. 45–75. DOI: 10.1017/S0269888913000349.

Zhou, Lu, Michelle Cheatham, and Pascal Hitzler (2019). "Towards Association Rule-Based Complex Ontology Alignment". In: *proc. 9th Joint International Semantic Technology Conference (JIST)*. Ed. by Xin Wang, Francesca Alessandra Lisi, Guohui Xiao, and Elena Botoeva. Springer, pp. 287–303.